HE BELL JOURNAL OF ECONOMICS
ND MANAGEMENT SCIENCE

Vol. 1 1970

Editor:
Paul W. MacAvoy, *Massachusetts Institute of Technology*

Associate Editors:
William F. Sharpe, *University of California at Irvine*
Harvey J. McMains, *American Telephone and Telegraph Company*

Managing Editor:
Gray L. Bromleigh, Jr., *American Telephone and Telegraph Company*

1

Cal -H05435- 1 - P80 23

CALCUTTA UNIVERSITY
CENTRAL LIBRARY

PRING '70

NO. 1.

Table of contents

3 **About this journal**
F. M. GARLINGHOUSE

5 **From the editor**
P. W. MacAVOY

6 **Regulation and the free market: the problem of boundaries**
JAMES W. McKIE

27 **Stopping rules for selling bonds**
WILLIAM M. BOYCE

54 **Regulation and the durability of goods**
RICHARD L. SCHMALENSEE

65 **Alternative rate of return concepts and their implications for utility regulation**
EZRA SOLOMON

82 **Some analytical results on tax depreciation**
PETER B. LINHART

113 **The theory of public utility pricing and its application**
RONALD H. COASE

129 **Resource allocation and the regulated firm**
ELIZABETH E. BAILEY AND JOHN C. MALONE

143 **The valuation of public utility equities**
BURTON G. MALKIEL

About this journal

A few years ago, a group of Bell System officers, in the course of a study of new concepts in public utility economics and finance, noted the need for a professional medium specifically devoted to fresh, scholarly thinking on these subjects. It seemed to us, and to those with whom we consulted, that the dissemination of objective, scholarly articles about advances in the economics of regulated industries would greatly assist the regulators and managers to progress toward their mutual goal of better serving public needs.

Since the initial idea was conceived, we felt that the field of management science also offered a body of knowledge that should be applied to regulatory issues. Consequently, we decided to combine these areas and to publish a journal whose purpose is to encourage and support research in the issues and problems of regulated industries. The result is THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE.

In publishing the BELL JOURNAL, the Bell System hopes to:

—expand scholarly interest—and creative thinking—in the application of economics, finance, and management science to the problem areas of the regulated sector.

—provide a forum for discussion of serious study in these areas.

—set forth for talented young economists and management scientists some challenging problems awaiting their interest and energy.

The name of the Bell System was explicitly and purposefully associated with the BELL JOURNAL with full knowledge of the risk that some might construe it as propaganda. However, we believe that the high quality of the articles themselves will attest to our intent. The value of the publication—and its ultimate success in stimulating research, thinking and understanding in the issues and problems of regulated industry—will be directly related to its continued intellectual independence and quality.

Thus, Paul MacAvoy of MIT accepted its editorship on the condition that he would select his own editorial board *and* establish the system of selection for papers submitted to the journal. The election of the one Bell System member to the board was based entirely on the individual's accomplishments in the management science field.

Every article in the BELL JOURNAL will be selected by Professor MacAvoy and his editorial board. The Bell System does not necessarily endorse the topics, approaches, conclusions or interpretations of any of the articles published. With some of the views expressed here we shall agree, with others disagree. That is not important. What matters is that the very best thinking be applied to the tasks and opportunites ahead.

F. Mark Garlinghouse
Vice President
American Telephone and
Telegraph Company

From the editor

You will find an unusual combination of subject matter areas in THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE. Whereas most professional journals limit their horizons narrowly enough to preclude combining economics, finance, and management science between the same covers, we invite papers from all three areas. Rather than limiting our interest to a single academic discipline, we seek articles in several disciplines.

The subject matter of this journal is the regulated sector of the economy. The BELL JOURNAL'S purpose is to support, encourage, and expand the outlets for new research on the regulated industries and, in particular, the large, regulated corporation.

Despite the growing importance of the regulated sector, very few articles appear each year on the economic theory of the regulated firm or the economic effects of regulation on industries and individual firms. Similarly, very little new work is published in the management science literature on the relations among scale, complexity of outputs, and regulation in the firm. There is, to be sure, a great deal of highly original work going on in management science, but relatively little of it focuses on the problems of management under regulation, and little of what does sees the light of publication.

The BELL JOURNAL invites both theoretical and applied research on the economics and finance of regulation. We invite articles on management science topics as they relate to regulated industries and firms, and on the general problems of control and optimization in large corporations. The only criterion for selection is whether, in the judgment of the Editor and Associate Editors, an article contributes materially to existing knowledge in economics, finance, or management science pertaining to regulated industries.

The Editorial Board undertakes to furnish the author of a submitted article with a decision on publication within a month of receiving the manuscript. In general, the editors comment on manuscripts freely and extensively, in the thought that such an active exchange can only benefit all concerned. A modest honorarium is extended to each author when his article is published. Accepted articles are normally published in the following issue.

Paul W. MacAvoy
Professor of Economics,
Sloan School of Management
Massachusetts Institute of Technology

Regulation and the free market: the problem of boundaries

James W. McKie
Professor of Economics
Vanderbilt University

The problem of setting appropriate limits to the direct regulation of public utilities is the subject of this article. The problem concerns both the internal and the external relationships of the public utility enterprise. Regulation cannot control all of the decisions and internal relationships of the public utility. Regulatory bodies therefore must economize the means and limit the scope of regulation; but they encounter considerable difficulty in doing so, because the regulated decisions affect the unregulated ones and the ultimate reactions may escape control.

External limits to regulated activities of public utilities must also be determined. A boundary problem appears whenever regulated activities are intermingled with non-regulated ones, or when public utilities compete with each other or with unregulated industries, or where technological innovation and new forms of organization lead to new kinds of activities for the public utility. Control authorities are always looking for neutral interfaces which will set boundaries to the regulated sectors without affecting activities outside them or relationships that cut across them. But such boundaries are not easy to find.

■ No human activity escapes regulation of some sort. In our economic life a minimum of government regulation is taken for granted. Though this minimum has tended to increase as economic organization has grown more complex, the free market is still the normal way of organizing economic activity. The conventional economic unit is the unregulated private business. Government regulation, especially of the detailed form commonly known as utility regulation, is the exception. In what follows we shall not explore the general virtues of the free market in relation to those of central control to find which one might be superior as a way of organizing a national economy. We shall merely assume what is essentially a value judgment: that direct regulation should be limited, if not minimized. If government regulation is applied to an activity or organization because the free market fails to work, it should be confined to those areas and decisions which competition cannot properly regulate. It should not be extended unnecessarily. The problem of drawing the appropriate boundary around the direct regulation of economic activity is the subject of this paper.

Public utility regulation at first glance looks like a fairly easy case. The class forms a relatively clear-cut exception to the normal

---

The author obtained the B. A. degree from the University of Texas in 1943 and the Ph.D. degree from Harvard University in 1952. He has published extensively on economic subjects and has served as consultant to business firms and government agencies on matters of regulatory policy. In July 1970, Professor McKie joins the Brookings Institution as Senior Research Fellow for a year of further study on the boundary problem in public utility regulation.

This article is an interim report on research sponsored by the Brookings Institution's Program of Studies in the Regulation of Economic Activity.

role of the free market, the grounds for the exception are fairly well known, and the policy is a long-standing one. Yet the boundary problem is there. It is becoming more difficult, and presenting new aspects for solution, as organization and technology develop and competitive activities press more closely upon the regulated ones. The problems fall into two main groups: external and internal to the regulated firm. The external ones arise in the effort to demarcate the regulated industry from the competitive sector and to determine the role competition will have in ordering its behavior. The internal ones arise largely in the effort to economize the means of control within the enterprise.

The internal limits of control

■ Let us glance at the internal problem before turning to the external boundary, which is the principal issue for this paper. Professional economists certainly realize how difficult and complex the task of control is even in organizations of small size and simple structure, which most public utilities are not. Managers know how much has to be left to routine, more or less automatic, indicators and rules, and impersonal mechanisms of control that simulate certain rudiments of a price system within the enterprise. The number of economic variables and policies that a business enterprise must determine somehow in running its internal affairs and its relationships with the outside world is formidable enough to discourage those who would attempt to control them all in order to make the enterprise conform to a predetermined plan. The following is only a summary list.

The price structure—minimum prices, maximum prices, the degree of discrimination

Non-price aspects and marketing methods

Competitive tactics

Efficiency of production; minimization of costs

Quality of service, in several dimensions

The rate of return

Investment and extension of service by type and area

Financial management

Innovation and choice of techniques

Purchasing policy

Employee relations—composition of the labor force, productivity, wages and other terms of employment, collective bargaining

Structural organization, acquisitions, integration, affiliation.

Control over these variables by the management of the enterprise itself is never complete, even with full access to information and complete internal authority. An outside agency like a public utility commission faces a far more difficult task—or would if it tried to control the organism completely. But a commission does not try to control it completely. Ideally, it tries to make the enterprise conform to a few simple performance criteria, leaving other matters to the discretion of its management.

No doubt the regulators' task would be simpler if there were an isolated mechanical correspondence between certain of the above-mentioned elements of business policy and the dimension of performance that the regulators wish to affect. If their primary purpose were to prevent excessive returns in a protected monopoly situation, a simple limit on rate of return would then be enough. Some complications arise because regulatory authorities actually apply a variety of welfare criteria, which do not always reconcile with each other. But the main source of complications is, of course, the interconnections among the decision variables themselves.

Possibly the rate-of-return standard gives less trouble in this respect than some other regulatory policies; that indeed is why it is so often chosen as a target variable. It represents the most fundamental determinant of the firm's economic performance, on which many other things depend. This approach, essentially one of using a single key variable to regulate a whole system of variables, will work best if all other variables whose behavior is material are linked ultimately to the regulated one in a kind of pyramided chain of dependent responses. Nothing could be more conducive to economizing the means of regulation than causing a small but easily grasped tail to wag a large dog.[1]

Spillover effects and compensatory variations among the decision variables can interfere with attempts to economize the means of control through rate-of-return regulation. The strongest of these compensatory effects is probably on productive efficiency, or the generation of "organizational slack." Once the rate of return as defined by the regulators is at its permitted maximum, managerial incentives to keep costs down are severely weakened, and excess expenditures on executive perquisites and corporate image are likely to appear.[2] Management pressure on the cost factors in production usually slackens. If regulators fix a rate of return designed to capture any increase in earnings above a minimum and do not allow any time lag in applying it, the management's incentive to introduce cost-reducing improvements also suffers. A closely allied reaction, when the allowed rate of return is above the cost of capital, is the so-called Averch-Johnson effect, or a tendency to overinvest and to favor unduly capital-using innovations and techniques.[3]

Any regulatory commission that tries to control these effects by regulating additional variables such as cost performance, executive salaries and perquisites, choice of technical methods and rates of innovation, will quickly find its hopes to economize the means of regulation evaporating. As it extends further into the network of enterprise decisions it may discover that still other compensatory changes partly frustrate its efforts, and there are always more just over the horizon. Extension of control in response to perpetually escaping effects of earlier regulation may be called the "tar-baby effect," since it usually enmeshes the regulatory authority in a control effort of increasing complexity with little gain in efficiency but a growing feeling of frustration.[4] I do not say that these induced variations must nullify any attempt to regulate the rate of return. But the values and relations of the other decision variables do not remain constant when an attempt is made to regulate one of them. As one is pushed close to an assumed optimum, some others may be pushed away from their optimum positions, while attempts to prevent this can lead to an indefinite extension of control.

---

[1] A parallel example of the key-variable technique outside the public utility field is the approach of the U S War Production Board in controlling economy-wide industrial production in World War II. It did this essentially by controlling use of just three metals: steel, copper, and aluminum Allocation of these three materials, which were in very wide use and in short supply, carried along most other wartime allocative decisions, enabling the authority to determine most of the composition of ultimate output without involving itself in a multitude of other decisions on input, output, and production. The key-variable approach had its drawbacks, and inevitably there were spillover effects The relation of the supplies of the three key metals to each other gave trouble, as did the special allocations of materials that were in even shorter supply than these three but with a narrower range of use Not all activities could be controlled this way, and inconsistency with other phases of regulation such as price controls gave trouble. Such a system would not work in the same way in a peacetime economy not as heavily dependent on strategic metals, but in the disequilibrium economy of wartime it did permit a remarkable leverage of control. See [8].

[2] See [16].

[3] See [1]. For a graphic exposition, see [14].

Multiplying the objectives of regulation, such as performance standards, further complicates the task. Suppose, for instance, that the regulators have certain criteria for maximum prices and for equitable discrimination. A simple ceiling on the rate of return which leaves the utility free to determine the rate structure as it pleases is likely to produce some very unsatisfactory results—extremely high prices in markets where elasticity of demand is extremely low, and degrees of discrimination that are likely to cause many customers to feel unfairly treated. If the authority fixes specific rates lower than the utility would have set them even under a rate-of-return constraint, however, the utility may well allow quality to deteriorate or even refuse to serve some markets at all. This can provoke the regulatory authority to control these variables also, thereby meeting itself coming around the other way from regulating the rate of return and imposing obligations on the franchised monopoly to serve the public without choosing its customers. The obligation to serve itself raises some interesting possibilities about strikes, interruption of service, and the labor negotiations that lead to such results.[5]

The complications involved in the quest for effective control over the internal decisions of regulated firms resemble to some extent those of setting external boundaries to regulation. But some of the external problems are novel—especially in recent years, as new forms of inter-firm and inter-activity relationships have emerged.

The regulated-sector boundary

■ Several problems have always been implicit in policies that demarcate a regulated sector from the unregulated economy. The most universal one is what has come to be known as the "second-best" problem, or the misallocations that may result when the regulated sector is made to follow rules of marginal-cost pricing and "efficient" output, while the unregulated sectors, which include a great amount of imperfect competition, are left free to misallocate resources accordingly. There appears to be no solution for this short of central direction of the economic system, which of course might generate diseconomies of its own. In any even, over-fulfillment of the norm of efficiency in the regulated sector might well be better than leaving it to exploit a natural monopoly, regardless of the presence of monopoly elsewhere.

---

[4] There appears in one of the Joel Chandler Harris stories an alien creature made of tar, which overcame an attacker by enmeshing him. The harder you struggle with this sticky creature, the more bound up you become.

[5] Obligations to avoid service interruptions also put constraints on technology and interconnection. For instance, the "electric power reliability bill," which resulted from some spectacular power failures, would require regional and interregional coordination of power generation and distribution. As a consequence, the private utilities ". . . would have to give outsiders a large hand in the planning and operation of their electric generating facilities—something that has been, up to now, a management function " (*The Wall Street Journal*, April 3, 1969, p. 7.)

Apart from the anomalies of the second-best, there are numerous transmissions of effects across the regulated boundary that may also be called "problems," though not all cry out for solution. Some of these effects stem from the nature of the utility as a monopoly (if indeed it is one). Some stem from the regulatory policies themselves, because effective control usually produces a degree of disequilibrium. And some are generated on the outside by competitive developments and are transmitted inward across the boundary. Any of these may give rise to external tar-baby effects.

For instance, a policy of regulation which imposes a ceiling price on a particular service or activity lower than the price that would make the demand equal the available supply (however monopolistic the price would be), and which fails to compel enlargement of supply to meet the demand at a lower price, will require some other method of rationing. Nonprice rationing may force extension of regulation to the rationing method, which can take the form of regulation of decisions by the customers of the regulated activity.[6] Examples of such extensions are readily found in societies more *dirigiste* than our own, but they always threaten whenever a disequilibrium is transmitted across a regulated boundary. If I read the signs correctly, they are just over the horizon in the natural gas industry. Since it is an increasing-cost industry and not a natural monopoly, the natural-gas producing industry cannot be made to discover and produce more to satisfy a larger demand at a lower price. If price ceilings are below the market equilibrium, supply must fall short of demand—a shortage that can only temporarily be alleviated by more rapid depletion of existing stocks. Eventually, the regulatory authority may have to decide on what ultimate uses shall have prior claim on the available supply, whether flat-glass production in Ohio is more important than house heating in Massachusetts, etc. End-use rationing is entirely possible if the pessimists are correct in their forecasts of costs and supplies, and if price ceilings are kept below long-run marginal cost.

If every economic activity were isolated and self-contained, our external boundary problem would be trivial. The easiest kind of frontier to draw and to police is an uninhabited one. It arises, and with it the likelihood of external tar-baby effects, because of vertical and horizontal relationships with other activities—most prominently in the following ways:

1. When there is competition among regulated firms, affecting either their main activities or fringe activities that are under the regulated tent.

---

[6] Note that the reverse case is symmetrical the tar-baby effects of price *supports* in lieu of price ceilings A support price on cotton that fails to limit supply leads to sales abroad at prices well below domestic support prices. This enables foreign producers of textiles to export to the U. S. or to third markets at costs well below those of domestic textile manufacturers Loss of markets then elicits a subsidy to U S. producers who export to these third markets and limits on imports of these low-cost cotton textiles to the U S One imagines that the next sticking point on the tar baby might be a direct refund of the equivalent of the support price to a U. S. buyer of shirts who can prove they were made of domestic cotton textiles.

2. When there is competition among consuming industries which are treated differently by the regulated firm.

3. When there is monopsony and discrimination against suppliers by the regulated firm.

(These last two types were referred to earlier. As noted there, regulation might attempt to deal with them *internally*, by controlling or altering the policies of the regulated firm, or *externally*, by extending controls to competition, prices, inputs, and outputs of the firms affected by these policies.)

4. When a regulated firm carries on activities that are unregulated but intermingled with its regulated activities.

5. When a regulated activity competes with an unregulated one.

6. When innovation or structural change creates new activities that may be eligible for regulation, whether or not carried on by presently regulated firms.

Marking the boundary of regulation in these circumstances is a delicate task. To draw it too far out causes inclusion of some activities that competition could manage as well, or perhaps better, and possibly suppresses independent supply of those activities. To draw it too close in would cut across some activities that are organically incapable of separation without loss of efficiency and might permit some monopoly functions to escape control. Once the boundary is drawn it must be guarded to prevent inappropriate crossing, or the appearance of new activities on one side that should be on the other. For this purpose, what the regulatory authority needs most are neutral "interfaces," or points of division in the continuum of economic activity that permit separation without appreciable diseconomies or irresolvable disfunctions on either side. Effective regulation without indefinite extension or external tar-baby effects requires identification of these planes of cleavage.

Neutral interfaces

■ Every firm includes a number of processes and activities that are linked together by organic connections and economic ties of one kind and another, which could not be organized and operated separately without loss of efficiency. That is basically what a "firm" must be when reduced to its minimum terms—a boundary within which it is more economical to organize and operate processes through central control of some kind rather than through the market.[7] The absence of neutral interfaces between processes is not unique to public utility enterprises. But the firm may include much more than the necessary minimum, transcending some logical boundaries which we should locate if we want to minimize the scope of regulation. Interfaces will not appear as long as substantial economies of integration of different processes are present. But even when they do appear within the total corpus of a utility enterprise, they are not much use for the purpose we are now discussing unless the economies of scale of the separable activities are less than those of the central productive processes of the public utility from which they are separated. If everything has to be a natural monopoly, the demarcation of boundaries would settle only jurisdictional questions among regulated entities. Hence we want to find neutral interfaces that also set boundaries to large economies of scale of the kind that create natural monopoly.

---

[7] See [3].

It is noteworthy that physical connection does not necessarily demonstrate economies of integration. Interfaces can appear at such points; the crucial question is whether there must be ownership integration, or the kind of central control and direction that might as well be ownership integration.[8] Nor do we want to prejudge the contrary proposition: central control and integration may be necessary, and no interface possible, even when the activity in question is not physically associated with the central production installations of the regulated enterprise. Such connections can arise because of risks and uncertainties, shifting intertemporal patterns of use, essential collateral services with an information barrier to outside enterprises, etc., though the proof of such economies of integration must sustain some heavy burdens.

Multiple uses of a utility's productive installation or plant equipment raise a parallel problem. Some of these multiple purposes may sustain the service mix of the utility, while other uses of the same equipment can supply outside markets that are not regulated. (An example is a communications network that uses computers for switching purposes, but finds that switching does not absorb the full capacity of a computer unit of economical size which can also be used for data processing.) Now if the outside market can be entirely supplied from the excess capacity of this equipment, it might as well be drawn under the regulatory tent. But if the market has room for a sizable number of outside installations which can operate along the same cost curve as the integrated facility, and can achieve unit costs as low at capacity use as the units used partly for utility purposes, the boundary should properly cut across these uses so as to leave the competitive one unregulated. If the equipment of the utility is used for both purposes, therefore, cost separations are usually necessary to permit it to supply both markets without compromising the regulation.

Cost separations

■ I realize that cost separation is highly unpopular in the economic literature; but that is because it is so often used to support full-cost or average-cost pricing, which is not what I am advocating here. Cost allocations are arbitrary, as everyone knows. They can be logical, appealing, and symmetrical, but still arbitrary, because they divide something that is in fact united. The common cost has to be allocated because of regulation, not because allocations reflect rational resource allocation. Nor should prices be based on fully allocated average costs when activities use common facilities. Prices should have marginal cost as a minimum, as economic analysis has always told us.

---

[8] William R. Hughes has shown how the market can in some circumstances effectively coordinate interconnection *among* regulated utilities without the need for central direction [5]. To use another example, the natural-gas producing industry is physically integrated with the gas pipeline-transmission network, yet it is often asserted that the field producing stage, having an essentially competitive structure, could be left unregulated, i.e., that a genuine neutral interface exists when production meets pipeline transmission, and the economies of scale that provoke regulation are all on the transmission side of the interface. In any case, integrated forms—gas companies that combine the functions of transmission and production—exist side by side with disintegrated forms. The "independent" producers and the gas pipelines having no producing properties are at least as efficient as the integrated ones, and are in fact the predominant form.

In fact, if all prices were always equal to marginal cost, there would be no problem of cost allocation. In that case prices would follow the best rule of resource allocation in each market, and only the overall rate of return would be influenced by average cost. In view of the relationship between marginal and average cost that actually prevails at realistic demand limits in most public utility markets, the rate of return would probably be negative, at least in the utility sector of the business using common facilities.

But, of course, prices are not set equal to marginal cost, either in the utility sector or out of it. Marginal costs are only a minimum at best. Utility enterprises are free to earn their regulated rate of return by raising some prices well above this minimum, up to the revenue optimum determined by the elasticity of demand (or sometimes to a lower ceiling fixed by the Commission). Its prices in the unregulated sector will depend on the state of competition there; it will presumably maximize the net contribution above out-of-pocket cost there. Within the regulated sector, most utilities have price structures that are discriminatory among uses and frequently among units of demand in any use or market. The marginal unit may have a price or average revenue equal to marginal cost, but the intra-marginal units will not. And if the utility can raise these intra-marginal prices (up to demand limits)—if it can increase the degree of intra-marginal discrimination, or carry off more of the consumers' surplus, when its rate base or cost basis increases under a fixed rate of return—then its course is clear. It will pull as much of the common costs of joint facilities into the rate base or within the utility cost boundary as it can—preferably all, as long as it does not run out of power to increase net revenues.[9]

Regulatory commissions have typically dealt with this threatened inflation of earnings in two ways. (1) They have required the utilities to add to their regulated revenues their earnings from non-regulated business to determine the overall rate of return on total investment including joint facilities. This might work in some cases, but it amounts to extending regulation of the rate of return into the unregulated sector as well.[10] (2) More commonly, commissions have required the utility to set up cost allocations at the boundary. If there are independent suppliers of the unregulated sector, this is at least a workable method of demarcation, and perhaps the only one. Its drawbacks, apart from its encouragement of average-cost pricing standards in both sectors, are those of any arbitrary division of an organic entity. These have long been familiar in cost allocations that provide demarcation between Federal and State jurisdictions, for example in telephone service, where they have led to disequilibrium and anomalies even though *both* sectors are regulated. Anomalies can occur if regulation on different sides of a boundary operates accordingly to different principles.[11] They can also occur if the allocations themselves force different pricing results on either side. When regulation is only on one side of the border, the unregulated sector is in theory unaffected by the cost allocations. But the regulated sector is heavily affected by them, and anomalies such as different prices for similar services are almost bound to occur.[12]

---

[9] This is recognizable as an aspect of the Averch-Johnson effect.

[10] Any cost allocation that distributes common costs or investments in proportion to earnings has similar effects. The early plans of the FPC to allocate joint petroleum costs among oil and gas in proportion to revenue would have led to further reallocations of cost as regulated prices for gas were lowered, until the joint return on total investment reached the regulated maximum, or until net gas revenues reached zero.

[11] A familiar consequence has been a lower rate for interstate long-distance calls (where rates are based more strictly on costs) than on intrastate calls of equal distance, in most states which have imposed rate-of-return control without strict rate ceilings on toll service.

Competition and the external boundary

■ The boundary between the regulated utility and the competitive sector is not inviolate nor immutable. Substitutes and alternatives for its services may appear. Indeed, they have never been entirely absent. If nothing else, utilities had to compete for the consumer's dollar, as any monopoly has to do. Sometimes they faced somewhat closer alternatives—telephone vs. telegraph, for example. During the last 40 years the growth of substitutes has accelerated. We have seen the rise of the motor truck and the revival of inland waterway transportation, the massive invasion of natural gas into fuel markets and the development of electric space heating, the appearance of new advanced forms of communication to provide alternatives to various parts of the telephone and telegraph networks, and many other developments offering challenges to regulatory policy.

Unfortunately, competition hardly ever replaces regulated monopoly. It usually invades some parts of a utility's service territory while leaving other parts untouched; it also takes some peculiar forms which do not permit simple substitution of a competitive market process for regulation of a natural monopoly. Hence the problem of workable boundaries presents itself. Let us consider several manifestations of that problem.

□ **Value-of-service pricing.** This is another name for discrimination, which an ideal public utility isolated from direct competition could practice perfectly within limits set by public control. With perfect discrimination, the monopoly would charge the demand price (the highest price consumers were willing to pay) for *each* unit of output down to the last unit, the price of which would be no higher than marginal cost. The rate of output would meet the economic test for an efficient use of resources: the demand price of the marginal unit of service produced would equal the incremental cost of producing it.[13] By discriminating, the utility would usually be able to earn enough to cover its total costs and make a profit, without having to "restrict" output as a simple monopolist would do.[14] Regulation could set bounds and limits to total earnings and to the degree of discrimination without disrupting the marginal conditions of allocation.

This at least is the approach that has been followed in the free economy to pricing by the privately owned regulated utility—though, of course, discrimination is far from perfect in any actual market. On this approach, value-of-service pricing above marginal cost permits the utility to meet the "total conditions." Unfortunately, competition tends to break it up. Things can go seriously wrong with discrimination when competition enters the picture; whether they do depends on the kind of impact it has and the kind of response the utility makes to it. The following discussion considers some illustrative situations producing different kinds of impact and response.

---

[12] Anomalies in the sense of different economic relationships on either side of the boundary are an inevitable by-product of regulation, even when it attempts to simulate competitive performance. But as long as the regulated sector is relatively self-contained, the anomalies in end markets merely take this form of the second-best problem referred to above. Common-cost allocations can produce much more visible ones For example, main-line pipeline sales of natural gas to industries are exempt from ceiling price regulation, so are intrastate sales. The "Atlantic Seaboard Formula" used by the FPC to allocate common costs between jurisdictional and non-jurisdictional sales is widely thought to require insufficient allocation to jurisdictional sales and, hence, a kind of "subsidy" of domestic consumers. See [4], pp 182–5.

[13] This result depends on the absence of significant income effects, i e., no appreciable shift of demand merely because the buyers paying higher intra-marginal prices suffer a small loss of real income thereby. It also depends on the fulfillment of other "marginal" conditions, which result in full cost efficiency in production, and on fulfillment of similar optimum conditions in all other economic activities.

[14] Cf [10].

1. Competition between regulated utilities and non-utility enterprises. The recent situation in the transportation industries exemplifies this kind of competition. Without attempting to settle several controversies in transportation policy, such as whether waterway transport receives a subsidy, whether user charges on highways are levied equitably against vehicles of various types, etc., let us simply assume that the private cost functions of these agencies are what they are and look at the new situation from the standpoint of the railroads and their regulators. Both forms of competition—first the waterway carriers, then the motor trucks—presented the railroads with new limits on markets and attainable prices.[15] Neither one had the large fixed-cost element represented by the rail roadbed. They had essentially competitive structures—small-scale units, large numbers, free entry.

If the competition had simply developed as a set of constant-cost supply curves for *all* agencies, the problem of adjustment would have been simple. But, of course, the railroads are not a constant-cost industry; they continue to have public-utility characteristics. Their long-run average costs decline with larger rates of output, as before. Indivisibilities, economies of scale, and the pressure toward price discrimination have not disappeared. What is gone is the opportunity to practice some kinds of discrimination, i.e., value-of-service pricing, as competitive services have appeared in those quarters.

Confronted with these competitive inroads, the regulatory authority could have tried several ways of establishing workable boundaries, short of abolishing regulation of transportation altogether. What it did do, as far as motor trucks were concerned, was try to make them behave like utilities. That is, it extended the regulatory boundary as far as possible into the competitive sector, ignoring the fact that a quite different kind of entity was now under the regulatory tent.[16] This is, of course, a standard kind of tar-baby reaction. The resulting anomalies in the transportation field have often been commented on. Motor truck lines and the conferences representing them were encouraged to adopt value-of-service discriminatory tariffs—something quite alien to their nature, which for that very reason soon began to erode toward a cost-of-service basis. As they did so, the persisting discriminatory tariffs of railroads caused shippers to direct traffic in the high rate brackets to truck lines, reinforcing the effect of an uncompensated quality differential.

---

[15] There were (and are) qualitative differences among the services offered by the three agencies, generally expressible as private service costs borne by the user These correspond to what Merrill Roberts calls "nontransport distribution costs." See "Transport Costs, Pricing and Regulation," in [7].

[16] It is not asserted here that the regulatory commissions themselves were solely responsible for this extension—it was a legislative policy. Yet all interests concerned with rail regulation seemed to move together toward regulation of the competitive challenge.

When the railroads perceived what was happening, they attempted to regain some of the lost traffic and hold what was left in the threatened brackets by rate reduction, down to marginal cost if necessary. Their rivals of course called this "retaliation." It led to an extension of the series of decisions and policy statements by the Interstate Commerce Commission, the Congress, and the courts, which had begun earlier in connection with rail-water competition and which obscured the economic issues behind a principle of fair shares for each agency of transportation.[17] The question of cost in relation to rates got transmuted into a "full-cost" or average-cost test for the division of traffic, especially as between those who had it and those who were trying to get it or regain it. The rail carriers generally were not permitted to capture, or regain, traffic from motor carriers by cutting rates below their own fully distributed cost if the rate had not been that low under monopoly conditions, even when the proposed rate exceeded their marginal or "out-of-pocket" costs and would have contributed something to overhead. (This same rule applied to trucks was not nearly so restrictive, since full costs or average unit costs were and are very close to long-run marginal costs for motor carriers.) Moreover, the ICC seldom if ever explicitly allowed for service advantages in fixing rate relationships. Marginal cost minimums for rates were not recognized as suitable even for rail-waterway competition. The "inherent-advantage" clause of the Transportation Act of 1940 was in fact a declaration of Congressional concern with equitable division of traffic rather than with efficiency. Its purpose was to protect competing carriers against regulation in the interests of the railroads.[18] The situation has improved since the Transportation Act of 1958 admonished the ICC not to hold the rates of a carrier up artificially to protect competitors, but it has not improved very much. The ICC and the Congress lack a reliable theory of pricing that would solve these problems.

In transportation, the question of the rate structure in a competitive environment is probably not extraordinarily difficult to solve. The long-run marginal cost of motor transport (virtually the same as its average cost) forms a schedule rising with distance, which intersects the corresponding marginal-distance-cost schedule of the railroad at differing points, depending on how much value the shipper attaches to the service advantages of trucks. (The railroad schedule rises much less steeply with distance, but the terminal costs are higher.) If all commodities and hauls for which its marginal cost is lower are to be served by the railroad, its rates must be correspondingly lower. This amounts to a marginal-cost minimum for price on most competitive traffic. At the other end of the service spectrum, the marginal cost of rail transport is below the marginal cost of water transport on some classes of bulk freight, especially when the service advantage of the rails is included. To get or keep this traffic, the railroad might also have to set rates near its marginal cost. The result would be a rational allocation of traffic among agencies, though it would not necessarily yield revenues to the railroad sufficient to cover total costs.

---

[17] See Transcontinental Cases of 1922, 74 ICC 48 (1922) and Tin Plate to Sacramento, 140 ICC 158 (1925) for early statements of this position, also "Regulation of Transport Industries," Second Report of the Coordinator of Transportation, Senate Document No. 152, 73rd Congress, 2nd Session (1934), and the Fourth Report, House Document No. 394, 76th Congress, 2nd Session (1936) The Transporation Act of 1940 set the tone for further application of such a policy approach. In the case of New Automobiles in Interstate Commerce, 259 ICC 475 (1945), the ICC declared against a policy of holding up the rates of one carrier to protect the rate structure of another mode, but it did not always follow this principle in its other decisions—at least, not to the extent of permitting a marginal-cost test for minimum competitive rates See Ingot Molds, Pa. to Steelton, Ky., 326 ICC 758 (1965) and subsequent court decisions The declarations in the Transportation Act of 1958 seem to convey a Congressional opinion in favor of competition, and the courts appeared to follow up this new approach in *New, York, New Haven and Hartford RR v U. S.*, 199 F. Supp. 163 (1961), but subsequent application of this doctrine has been erratic, as evidenced in the Ingot Molds case.

[18] E W Williams, Jr., asserts: "Indeed, the Commission has been moved seldom, if at all, by the conclusion that regulation has consistently, over the twenty years since the Motor Carrier Act, deprived the low-cost carrier of its cost advantage, a result often tantamount to depriving it of all opportunity to compete for traffic." See [15], p. 212.

The competing agencies would divert some traffic that used to belong to the railroads and would put ceilings on the rail rates for a wide variety of traffic in which rails used to have a monopoly. But a regulatory policy which recognized these structural facts could still sanction rate discrimination when it was possible, and even permit increases where the traffic would bear them. The regulatory authority would concern itself primarily with enforcing minimum rate standards, to ensure that the more powerful carriers did not set rates below marginal cost as a predatory tactic, as well as enforcing certain peripheral restraints on discrimination such as the long-and-short-haul clause. This mixed cost-of-service and value-of-service pricing system would probably leave the rails in a deficient-revenue position, necessitating a subsidy, which would be a break with past standards in the U. S. that have generally included a requirement that privately owned utilities support themselves. But fitting the railroad into a competitive environment might require it. It is certainly a more rational policy than attempting to make the competitive industries behave like utilities.

2. Competition Between Utilities. More troublesome problems arise when utilities themselves compete with each other. By "utility" I mean a firm having the typical cost and service characteristics of a utility, even though it does not have an unchallenged natural monopoly in all its services. There are some instances of rivalry under similar conditions in manufacturing industry, but I shall concentrate on the rivalry that involves regulated utilities in order to develop the boundary problem that faces the regulators. In general, the utilities in this situation have some monopoly services and some that encounter substitute competition from other utilities. There may also be competitive industries in the picture supplying substitute goods or services, as in the previous case.[19]

The problem that rivalry among utilities presents for regulation is not new: it appeared almost immediately after regulation began, in the form of terminal competition among railroads. In fact the price wars that broke out at those points helped to bring on regulation, and they certainly made the railroads more willing to accept it. Rates between major terminals served by parallel roads fell precipitously during rate wars of the 1870's and 1880's in trunk-line territory, while the roads maintained rates from intermediate monopoly points at levels that would exact maximum revenues from shippers who had no alternative. The low through rates were responses to a combative challenge, not a rational application of marginal-cost pricing. After Federal regulation began in 1887 the ICC prohibited secret reductions of rates and restrained price warfare. In some cases, it sanctioned rates determined by cartel agreements among railroads, though it forbade pooling of freight revenues.[20] Regulation also required abandonment of some forms of geographical discrimination that had resulted from terminal rivalry, specifically short-haul discrimination against monopoly points. At the same time the regulatory commissions had to recognize the facts of geography and permit different roads to charge the same rate between the same terminals regardless of relative route distance, to equalize rates into major markets and to principal river crossings, etc. They also sanctioned advanced forms of commodity and class rate discrimination.[21] Within these control boundaries, competition or rivalry among railroads was permitted to exist, principally in the form of service competition, but was not allowed to break out into price rivalry again until competitive agencies of transportation began to break down the controlled structure of rail rates.

---

[19] Rivalry among gas utilities, electric utilities, and dealers in fuel oil for space heating is an illustrative instance.

Rivalry among railroads between terminals in the absence of regulatory control is likely to be drastic—an extreme case illustrated by our most unstable duopoly models. Another utility confrontation is now developing, which is perhaps less drastic but carries with it some of the same problems. This is rivalry between electric and gas utilities for home heating, cooking, and water heating, and for certain commercial energy markets. Neither gas nor electricity for residential and commercial use is new. In the past, however, there was more or less a division of fields along a boundary of product preferences at the relative prices that obtained then. Space heating, apart from the declining share held by coal, went to gas—or to oil, near tidewater. Where gas dominated space heating it also dominated water heating and kitchen range use. Electricity had a monopoly on lighting and power for appliances. Regulatory policy had less difficulty with competitive relationships during this period.

But things never stand still. The electric utilities and appliance manufacturers generally improved the technology of utilization in fields competitive with gas. (The gas companies made some progress in uses competitive with electricity, though it was much less impressive.) The real cost of electricity declined relative to that of gas over a long period. The first effect of these was the development of a competitive challenge in range fuel and water heating—partly in the form of aggressive product promotion, but partly in the form of open or disguised rate cutting by electric utilities. The rivalry of electricity with gas for home and commercial uses received a mighty boost from the rapid growth of air conditioning in the postwar period. It quickly became a very large residential/commercial use of electricity, and in most areas it caused a reversal of peak demand load from winter to summer. This was a momentous event for the utilities. It impelled them to take a new look at gas-competitive uses in the home, principally space heating.

The outcome has been unsettling both to the utilities and to their regulators. Rivalry has found its way into a kind of disguised price warfare, and the standards for price control have become very uncertain. The rivalry for range fuel for at least 20 years has often taken the form of promotional allowances to the consumer, or to the builder or developer who installs facilities and appliances in the housing he sells or rents. These promotional payments became much more significant—even news-worthy—when the electrical utilities began to invade the residential heating market. Perhaps everyone has heard about the promotional payments and allowances of $1,500,000 to the builders of the Chicago John Hancock center.[22] These payments were to ensure use of all-electric energy, including space heating, in this giant apartment house. There have been many other well-publicized instances, as well as many less spectacular promotional payments by gas companies. Some of these have been secret, individually negotiated, *ad hoc* payments, while others have been openly announced and available to everyone who would use electricity instead of gas, or gas instead of electricity, for specified types of consumption.

---

[20] See [6], Ch. 5.

[21] Ibid., p. 113.

Whether promotional allowances and payments made to builders are or are not passed along to home buyers and renters is a thorny question that I would prefer to bypass. The more important issue is the one they pose for rate-making and regulation. This in turn breaks down into two questions: (1) Should promotional payments be controlled on the same basis as the rates themselves? (2) What standards of control are appropriate for utility rates in a competitive environment? In my own view, with which not everyone would agree, allowances and payments are equivalent to rate adjustments and should depend on the same standards. This enables us to pass at once to the question of what those standards are or should be under the circumstances that prevail in those markets.

The reaction of some interested parties to the appearance of rivalry in local-utility markets has been much the same as that which rivalry among transportation agencies elicited: a full-cost standard, with market divisions governed by equity, or at least protection of entrenched interests from inroads by substitutes. For example, Mr. W. L. Lee of the Atlanta Gas Light Co. recently asserted:

> In no event should a practice [i.e., a promotion to secure new business for a utility] be permitted unless the net revenue, on a fully distributed cost basis, derived from the additional load promoted, recovers the cost of the promotional allowance. Otherwise, the remaining ratepayers of the utility will be forced to subsidize the allowance. Today's customers should not be required to subsidize those of tomorrow. A fully distributive cost basis, advocated for many, many years by most commissions, eliminates most of the hanky panky in rate design.[23]

Economists usually experience a feeling of profound depression when they hear opinions like these. Marginal cost, we say, is the correct test, not fully distributed or average cost. If a service yields revenues in excess of the marginal cost of supplying it, it benefits the firm offering it. If incremental revenue is not less than incremental cost, the activity does not have to be "subsidized" by other customers, either those of today or those of yesterday. There is nothing to be recouped; the activity in fact adds to the net amount available to cover overhead and net profit. Hence it benefits the other customers by reducing their burdens. Economists have enunciated these principles on many occasions,[24] and I am certainly not going to deny their essential correctness. I do hope, however, to reveal or clarify certain difficulties in their application to the problem at hand—that of determining the proper role for competition among regulated utilities and the proper basis for price (rate) regulation in a context of rivalry.

---

[22] See [18], pp. 537 ff.

[23] Ibid., p. 564. See also [17].

[24] See, for instance, [2].

☐ **The marginal cost test.** Economic theory contemplates a range of marginal cost functions, from the shortest of short runs to an ultimate complete and timeless variability of factors. It holds that short-run marginal cost is the test of efficient price.[25] When this test is applied to a complete utility system, short-run and long-run marginal cost tests for price come to the same thing over time. What is required for this equivalence is correct anticipation of the future. The long-run marginal cost test simply determines the optimal adjustment of capacity.[26] But practical problems arise when the total demand is broken into a multitude of sub-markets in an imperfectly discriminatory pricing system. If rivalry prevails, and if customers have different degrees of bargaining power, the marginal-cost price for the whole system can be obscured by a spurious ordering of customers at the margin of consumption. The long-run incremental costs of expansion can easily be lost from sight beyond the immediate exigencies of the pricing situation. Reasonable customer classifications and a rule against personal discrimination are regulatory defenses against this. The regulators forego the benefits of moment-to-moment changes in rates, reflecting changes in short-run marginal cost, in order to secure the benefits of published tariffs and advance commitments on rates.

Without such safeguards, a utility might well attempt to justify extreme cases of rampaging rate war by arbitrarily selecting a "run" short enough and a marginal cost low enough for some bits and pieces of disputed business—at the extreme a marginal cost for each buyer, in the inverse order of their bargaining advantage. The cutthroat rates among railroads in the 1870's probably met a short-run marginal cost test of some sort, in the sense that a concession to a particular shipper at least covered the extra fuel and labor costs, even though other shippers of the same commodity over the same line at the same time didn't get it.

There seems to be a cautious agreement that short-run marginal cost divorced from capacity expansion is a proper test for rates if there is a marked load-factor advantage involved in the decision. Where capacity is not fully utilized, short-run marginal costs are economic minima for off-peak prices. But in our electricity-gas case, for example, it is one thing to apply such a minimum to a water-heating rate when the controls on the heater switch it off during the peak periods. It is another to apply it to a space-heating rate on the assumption that demand will absorb only excess capacity, when a successful generation of additional massive demand could move the utility peak back to the winter, presumably piling all of the future facilities costs onto the heating rate and taking them off the air conditioning customer. (The TVA experience shows that this is entirely possible.) Perhaps only a time variation rather than a type-of-use variation in rate levels is appropriate for load factor adjustments.

---

[25] F. M Westfield has argued that when this principle is applied to utility pricing, "Marginal-cost pricing requires that at each moment in time, at each location, price be equal to marginal cost of electricity delivered at that moment. . . . Just as the price of fresh strawberries in ideal competitive markets (which marginal-cost pricing is to simulate) must at each location be forever changing so as to clear markets at the point of equality of short-run marginal cost with price, so the rule for the price of electricity must aim to equate at each location short-run marginal cost with price at each moment of time." [13], p. 69. The marginal cost test does not work satisfactorily if the cost function is not continuous.

[26] Ibid. See also [12], pp. 91–2. The economic implications of this equivalence seem to have been worked out first by M. Boiteux and his fellow architects of the French *Tarif Vert*.

The marginal revenue also depends on how long the price or rate in question prevails and how customers and services are classified. It can be very volatile under flexible pricing policy (such as the famous "midnight tariffs" quoted by rival railroads to favored shippers) or when utilities can offer bribes to particular customers. The incentive to do this grows whenever rivalry intensifies. The results, if uncontrolled, are far different from those of regulated utility tariffs when personal discrimination is prohibited. I don't say that such a pricing system would be impossible. Some might even say that it would be optimal. From the customers' point of view it could begin to resemble bargaining in the bazaars. Opening one's monthly electric bill would be a real adventure. But such a system of ratemaking seems incompatible with public regulation. Nor would it really simulate competition, in the structural circumstances that prevail.

☐ **The direction of choice.** To serve the ends of allocation of scarce resources, utility pricing methods when substitutes compete should induce consumers to choose the service with lowest marginal cost—or, if there is a service advantage, to pay a price which reflects the higher marginal cost of a service with advantages that may offset the higher cost. In other words, they should not be led by the vagaries of utility pricing to consume the high-cost service when they would consume the low-cost service if each were priced at its marginal cost. Yet this can easily happen if competitive undercutting of rates takes place and "retaliation" is prohibited, as we have already observed in connection with truck-rail competition.

Critics of utility regulation have pointed out that "internal subsidization" by utilities invites such results. An economist would interpret an internal subsidy to mean a price for some service below the appropriate marginal cost, resulting in a loss that is recouped from high-value services. Richard Posner has put it this way:

> [I]nternal subsidies promote misallocation of resources. An inefficient competitor may be attracted to a regulated market by the fact that the monopolist is maintaining so large a spread between his costs and his rates in that market (in order to subsidize below-cost service elsewhere) that the entrant can make a profit even though his cost of service is higher than the incumbent's. . . . In these circumstances, price cannot fulfill its function of directing resources to the area where they can be employed most efficiently; rather, it invites wasteful duplication.[9]

An economist can certainly agree with this criticism; all that Mr. Posner fails to see is that it applies with equal force to any value-of-service pricing system that comes under competitive fire from a rival utility, or even from a substitute industry. It is not necessary to have internal "subsidization," i.e., rates below marginal cost, for this to happen. Any rate above a rival's marginal cost is a natural target.

P-5023

☐ **The workability of competititon among utilities** What is the answer to this problem? It is tempting to say that rates in competitive situations should fall to the relevant marginal cost, which provides an optimum answer to the allocation question. The trouble with this solution is that it does not guarantee that the utility will be able to meet the "total conditions" by covering its costs. It would not be able to if all rates fell to marginal cost under competitive pressure and if marginal cost were less than average costs, in the long and short runs.[27] Not everyone shares my pessimism. of course. Irwin Stelzer, for example, recently offered this prescription for utility pricing in a competitive environment:

> An appropriate rate should, at a *minimum*, be in excess of the complete incremental cost involved in providing the services. . But the lower limit of such a competitive rate—the complete incremental cost of providing the service—is not necessarily the most appropriate level at which to set that rate. . . Rather, at that point *or at some point above it*, there is a rate level which will maximize the incremental net revenues from that service and which is therefore the best rate from the point of view of all customers. At that rate this new business will be making its maximum contribution toward overheads. The precise rate satisfying this condition will depend on just how sensitive demand is to price: the more sensitive, or elastic, demand is with respect to price, the lower will be the optimum rate.[28]

This assumes in effect that there will be enough services left with low enough elasticities to permit the utility to earn its necessary rate of return; it also assumes that the elasticities on the competitive services will be stable and calculable. If a competitive, constant-cost industry offers a substitute service, the demand for the utility service may indeed become perfectly elastic (permanently) at that price, which constitutes a ceiling, but when the substitute is offered by another utility it isn't so simple. The pricing ranges overlap. Unless prices are equal to a stable marginal cost for all units of the competitive services, therefore, the "competitive" price relationship can be uncertain, unstable, and interdependent. The duopoly model again comes to mind: the "elasticity of demand" for each duopolist is not calculable in the face of independent pricing by the other. The marginal revenue is unknown. There is no equilibrium. At its worst, rivalry under such conditions becomes price warfare and can degenerate to mere blind bashing by infuriated antagonists. If anything, the cost conditions prevalent in utility competition only aggravate this result, at least in the absence of a regulatory check.

Some observers speak of competition among utilities as if it were a sufficient guarantee of acceptable performance; as if the mere appearance of a substitute service created workable competition and made rate regulation unnecessary. They probably would not accept such a conclusion about competitive behavior for an industrial duopoly such as plate glass, nor perhaps for direct competition between two electric utilities or two railroads in the same market. I cannot see that substitute competition guarantees a workable result either, nor that the regulated utility can be returned to the competitive market sector as soon as a single substitute appears on the horizon if regulation of the "natural monopoly" was appropriate in the first place. Rate warfare is not competitive equilibrium. Of course, if the competing utilities do remain under regulation, the commissions need to know how to control them and how to admit substitute competition into the regulatory scheme without producing chaos. I am not sure that economists can tell them how to solve all the resulting problems, but we can at least advise them to avoid fully distributed cost as the criterion for competitive prices.

---

[27] It has sometimes been suggested that the proper solution to this problem is to charge every customer a lump sum for the privilege of using the utility's service—a fee that would not affect the rate of use Variable consumption could then be priced at marginal cost, and the fee would make up the difference between the variable and the total costs. (I am indebted to Professor P. W. MacAvoy for calling my attention to this) Such a solution is sound in principle and might even be workable in practice, at least in some cases If the fee were made to vary among consumers, of course, there would be little difference between this policy and perfect discrimination if neither generated any income effects upon consumption.

[28] See [11], p. 12.

Let me return briefly to the promotional allowances by gas and electric utilities that I mentioned earlier. These offer some good examples of how rates and terms can get out of control under conditions of rivalry among utilities. What is at issue here is not "competition" at all—everyone feels instinctive approval upon hearing that word—but the degrees and kind of discrimination. The utility offering these promotions has no intention of giving up its franchise protection, nor of adopting a general uniform competitive price, nor of abandoning its remaining monopoly service. It simply wants to quote lower effective prices on the particular fragment of business that it is trying to wrest away from a rival. Again let me quote Stelzer: "Because new customers are, for electric utilities, the focal point of the fiercest competition with other fuels . . . is it not, theoretically at least, appropriate to treat these new customers as a separate class for rate-making purposes. . .?"[29]

We know what would happen if Goodyear Tire or Standard Oil did this. In the utility field it can lead not only to highly volatile rates but to widespread personal discrimination, as we have seen. Perhaps stable rates and reasonable customer classifications would no longer be necessary under conditions of rivalry, but I still doubt whether "competition" can take over the whole job of utility regulation in the public interest, even when reinforced by the antitrust laws.

☐ **Cream-skimming.** One more implication of the new competitive developments should be mentioned: the question of "cream-skimming," or entry by competitors into those markets where prices are considerably above marginal costs. One form of this was just described, as when a utility invades another utility service whose rate is above its own marginal cost. Other enterprises may enter such markets and attempt to avoid becoming full-fledged utilities themselves. The very phrase implies that regulatory commissions ought to forbid this, on the ground that competitors who go after the creamy markets and ignore the skim-milk ones upset the balance of internal discrimination or "subsidy" by the full-fledged utility and make it impossible for it to serve the skim-milk markets also. Or, in other words, Microwave Communications, Inc., should not be allowed to enter the lucrative Chicago-St. Louis trunkline market for microwave transmissions, because it does not plan to extend communication service to East Overshoe, Illinois, whereas A.T.&T., its target, does, and must continue that service as well as the trunk-line service which "subsidizes" it.

---

[29] [11], p. 13. Stelzer goes on to say that the rates of *old* customers should be lowered to the new-customer level after a period, e.g., 5 years. He also prefers lump-sum promotional allowances to this kind of new-customer rate reduction. (I have already stated my belief that these are equivalents.)

Again, economic analysts deny that there is any issue of "cream skimming"—that it is based on a false model of full-cost allocations. As long as the rate to East Overshoe covers marginal cost (as it always should), there is no internal "subsidy." Entry of new competitive services in the high-rated markets can only benefit consumers there by creating alternatives and (probably) lowering rates. And so on. Analytically, this is correct. It simply doesn't tell us what to do about the "total" revenue cost constraint. If there are remaining monopoly services that pay less than the traffic will bear, their rates will be jacked up when some cream is skimmed off elsewhere in the system. If there is no unmilked cream elsewhere in the system, the earnings on the whole utility system might fall below the necessary or allowable rate of return—and eventually the service to East Overshoe might have to be subsidized or else be lost with the demise of the system. (Of course, we would have to decide what the marginal cost of this service actually is under a set of contingent variations.) Either solution involves some difficult distributive, regulatory, political questions. To this extent the cream-skimming problem is genuine. I do not for a moment suggest that its solution is the suppression of substitute competition, but the present plight of the railroads shows how serious the problem can be and how difficult it is for regulators to find a solution based on their traditional approach to regulation.

☐ **Summary: utility pricing in a competitive environment.** Let us recapitulate the cardinal points of this discussion of utility pricing and the external boundary in a competitive environment.

As long as the utility is relatively isolated and self-contained, its pricing problems are relatively simple. A system of price discrimination or "value-of-service" pricing, subject to overall rate-of-return constraints, can be controlled as an organic whole by the regulatory authority, which can fix discriminatory differentials and demarcate rate classes in accordance with the economic and non-economic objectives that the utility is supposed to serve. The system is satisfactory from the point of view of economic welfare as long as it fulfills the "marginal" conditions.

The mere appearance of substitute "competition" need not lead to great difficulties as long as the resulting elasticities are stable and calculable and the impact does not invert or destroy logical or necessary pricing relationships within the utility's markets. But it *may* cause difficulties for regulation, especially if the rival is another utility.

(1) Competition may drastically affect the attainable degrees of discrimination, radically change the internal structure of profit margins and overhead coverage, break up customer classifications, create problem areas in the form of below-cost pricing of certain services, and raise some difficult issues of distribution of benefits and costs among users. Of course this can happen in any industry, but the effects are magnified in the typical utility firm, while adjustment through changes in number of firms and their products and markets is often impossible.

(2) Rivalry affects the stability of prices. It may lead to oligopolistic warfare if regulation fails to control the impact.

(3) Competitive erosion of a utility price structure can reduce its earning power below the regulated, and required, rate of return; there is no necessary relation between the utility price structure (under competition or otherwise) and a given rate of return. Meeting the "total conditions" is providential. Utilities in the monopoly state can usually meet them only because the potential earning power of an unregulated discriminating monopoly is usually very high.

(4) If relative prices fail to adjust themselves speedily to the competitive situation, because of regulatory lag or otherwise, consumers may be led to use the service with a higher marginal cost but a lower price than another. Yet the "correct" adjustment may be obscure and difficult to attain as long as value-of-service elements persist in the rates. The mere existence of competition forces some kind of rate adjustment as a rule and possibly even a general adjustment down to marginal cost.

It may be that the only rational pricing scheme in such circumstances is plenary marginal-cost pricing—that is, prices for all units equal to marginal cost. Some economists argue that this system is optimal and certainly better than a privately administered value-of-service system, even though it might require abandonment of the self-support principle and necessitate a subsidy or perhaps public ownership. When one contemplates the jerry-built structure of rates that results when utilities are attempting to find stable and remunerative prices in an environment that includes some substitute competition, some overt rivalry from other utilities, and some remaining monopoly markets, one may well yearn for a simpler and purer rule for price control. I will not go into the corresponding disadvantages, though we could of course invert the problem and talk about the Post Office and its impending escape from subsidy and public ownership.

In any case it is not easy to fit the total utility enterprise, with all its internal complementarities and massive economies of scale, into such an environment. The utility must be operated as a whole, and application of *ceteris paribus* rules in particular segments of its markets does not help to solve its overall problems. The difficulties we have noted—with cream-skimming, recoupment, the degree of discrimination, and potential price warfare—are all symptoms of the boundary problem among other things. They are no doubt capable of partial solution. But they remain formidable enough to warn us that we can't simply erase the boundary and incorporate the utility into the competitive sector as soon as competitors of one kind or another assault the frontier.

References

1. Averch, H. and Johnson, L. "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, Vol. III, No. 5 (December 1962).
2. Baumol, W. J., et al. "The Role of Cost in the Minimum Pricing of Railroad Services," *The Journal of Business* (October 1962), pp. 1–10.
3. Coase, R. H. "The Nature of the Firm," Economica, N. S., Vol. IV (1937), reprinted in American Economic Association, *Readings in Price Theory* (Chicago: Richard D. Irwin, Inc., 1952), 331–351.
4. Garfield, P. J. and Lovejoy, W. F. *Public Utility Economics.* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964.

5. Hughes, W. R. "Short-Run Efficiency and the Organization of the Electric Power Industry," *Quarterly Journal of Economics*, Vol. LXXVI, No. 4 (November 1962), pp. 592–612.
6. MacAvoy, P. W. *The Economic Effects of Regulation.* Cambridge: The MIT Press, 1965.
7. Meyer, J., ed. *Transportation Economics.* New York: National Bureau of Economic Research, 1965.
8. Novick, D., Anshen, M. and Truppner, W. *Wartime Production Controls.* New York: Columbia University Press, 1948.
9. Posner, R. A. "Monopoly and Regulation," 21 *Stanford Law Review* 609 (February 1969).
10. Ruggles, N. "Recent Developments in the Theory of Marginal Cost Pricing," *Review of Economic Studies*, Vol. XVII (1949–50).
11. Stelzer, I. M. "Impact of Competition on Regulation: Utility Rate-Making," Address to a Conference on Public Utility Valuation and the Rate-Making Process, Iowa State University, Ames, Iowa, April 25, 1968.
12. Turvey, R. *Optimal Pricing and Investment in Electricity Supply.* Cambridge: The MIT Press, 1968.
13. Westfield, F. M. "Practicing Marginal-Cost Pricing—A Review," *The Journal of Business* (January 1966).
14. Westfield, F. M. "Regulation and Conspiracy," *American Economic Review*, Vol. LV, No. 3 (June 1965).
15. Williams, E. W., Jr. *The Regulation of Rail-Motor Rate Competition.* New York: Harper & Brothers, 1958.
16. Williamson, O. "Managerial Discretion and Business Behavior," *American Economic Review*, Vol. LIII, No. 5 (December 1963).
17. The "Doyle" Report, *National Transportation Policy*, Committee Print, Committee on Interstate and Foreign Commerce, U. S. Senate, 87th Congress, 1st Session, 1961.
18. *Promotional Practices by Public Utilities and Their Impact Upon Small Business*, Hearings Before the Subcommittee on Activities of Regulatory Agencies of the Select Committee on Small Business, U. S. House of Representatives, 90th Congress, 2nd Session, July 16, 1968.

Stopping rules for selling bonds

William M. Boyce
Mathematics Research Center
Bell Telephone Laboratories

We model the problem of selling or issuing bonds so as to maximize the selling price (minimize the interest rate) as an optimal stopping problem for a random process or time series. Available information on, or predictions of, the price at a future time are included as a constraint on the process. Both continuous-time and discrete models are analyzed. For the case in which the predicted future price is normal or Gaussian we obtain good estimates of the optimal stopping strategy and expected gain. A significant conclusion is that the nature of the optimal strategy can be very sensitive to the relative variance of the predicted future price.

1. Introduction

■ We take our motivation from the following two bond-selling problems:

(1) An investor owns \$10,000 (face value) of bonds which mature in three months. The current market value is \$10,100. Should he sell now, sell later, or hold to maturity?
(2) A corporation must repay \$10 million in bank loans in three months, and it wishes to sell bonds to repay the loan. However, the company's economists predict that in three months bond prices will be lower (interest rates higher). Should the corporation issue the bonds now, wait a month or two, or wait the full three months?

These two problems share the elements of short-term price fluctuation and medium-term price information. In each, action must be taken at or before a fixed future time (or FFT), but waiting until the FFT will probably leave the seller worse off than acting immediately. The problems differ in the certainty of the FFT price; in the first, the exact maturity value is known, but the second seller has only an expert opinion, which might be represented as a probability distribution on the FFT price. Each bond seller's objective is to find a selling rule or strategy which will provide him a maximum (positive) expected gain despite the downtrend in prices.

In this paper we consider mathematical models or "games" which have the same elements of short-term value fluctuation, information on the values at an FFT, and the necessity of stopping (selling) at or before the FFT and taking the current value. Thus our

---

The author has been with Bell Telephone Laboratories since 1967, when he obtained the Ph D degree in mathematics from Tulane University. He also holds the B A and M.S. degrees in mathematics, awarded by Florida State University in 1959 and 1960, respectively. His current work is with stochastic models in economics and finance.

The author is deeply indebted to L. A. Shepp for introducing him to this class of problems and providing continuing advice and encouragement. The extent of his contributions can be readily judged by the reader. C. L Mallows and B. F. Logan provided assistance as noted F W Sinden originally suggested a maturing bond as the ideal security to which Shepp's Theorem 1 applied. D P. Heyman has provided helpful comments and suggestions. The help of M J. Hooper with the programming is greatly appreciated.

games will form a class of "stopping rule problems" [3]. In our games the "player" may face the same problem as the two bond sellers, that of trying to exploit the short-term fluctuation for an expected gain while facing certain (or expected) loss if he continues to the final time. We believe that our conclusions shed light on the analogous economic problems.

In section 2 we describe some assumptions of our models, then in section 3 we discuss our first model, the continuous-time or Wiener-process model. For this model L. A. Shepp has solved the optimal stopping problem for the case in which the FFT value is known exactly [8], which is analogous to the first bond seller's problem. Shepp also has partial results on the important case of the FFT price distribution being Gaussian or normal. In section 4 we introduce a discrete "urn" model, which is better suited to numerical calculations. In the next section we analyze the original model of this type, "Shepp's urn problem," which again represents the case in which the FFT value is known exactly. In section 6 we expand Shepp's problem to the "random urn problem," which models the general case of a probability distribution of the FFT price. We are able to derive an efficient algorithm for computing optimal strategies and expected values for an arbitrary distribution of FFT values.

Using this algorithm, we apply numerical calculations (section 7) together with theoretical arguments (section 8) in an attempt to generalize Shepp's solution of the "known-FFT-value" case to the case in which the FFT-value distribution is Gaussian. The results show that for virtually all standard deviations σ, a positive expected value can be obtained for games having a wide range of expected loss at the FFT ($\mu < 0$, where μ is the mean of the FFT distribution). Particularly surprising is the manner in which the type of optimal strategy varies with σ. For σ near zero, it naturally resembles Shepp's "problem 1" strategy, but above a critical value of σ, the type of optimal strategy is completely reversed. This example is significant in that it illustrates the danger of using a nonrandom model to derive policies for use in uncertain real-life situations.

In section 9 we work out an example based on our conclusions in sections 7 and 8. In section 10 we suggest an application to a stock-market phenomenon; and the last sections give closing comments.

Proofs are generally relegated to the appendices or omitted, since most of them appear elsewhere [2]. The appendices also contain some additional information on the urn models, which was not relevant to the main conclusions of the paper.

2. Modeling considerations

■ A complete analysis of a problem of the type we have posed should consider the effect of discounting (or the time value of money), the seller's utility function, and the modeling of the time series of price or value. However, we believe that the price model is most important for our conclusions. The influence of an explicit discount rate is minimized by the presence of a relatively close horizon (the FFT), and our gains and losses will seldom be spectacular or catastrophic enough to make the nonlinearity of utility a major concern. Thus we shall concentrate on the analysis of the effects of the price model and leave the inclusion of discounting and nonlinear utility to others or to another time.

(We should note, however, that the inclusion of discounting in our principal algorithm is simple, and the addition of a utility function requires only a minor reformulation. Taking "log-price" rather than price as the basic time series can be handled like a utility function.)

For modeling the time series of prices, we adopt the well-known hypothesis, first proposed by Bachelier in 1900 [1], that the market price of a security over time has the properties of a cumulative sum of independent random price changes—that is, a "random walk." Thus if $x(0)$ is the present price of a security and $x(t)$ its price t units in the future, then $\delta x(0,t) = x(t) - x(0)$ is independent of $x(0)$ and all earlier prices, and all the effects on $x(t)$ of prices up through time zero is exerted through the additive influence of $x(0)$,

$$x(t) = \delta x(0,t) + x(0)\,.$$

The "random walk hypothesis" does not deny that knowledge other than that of past prices may be useful in forecasting future prices, so it is reasonable for us to assume for "problem 2" that a forecast based on economic analysis has predictive value. Also, random-walk proponents often assume that the random walk is superimposed on a relatively long-term trend [6], and many of the cases we consider may be represented as a random walk plus a (known or unknown) drift or trend.

3. The Wiener-Process Model

■ We will consider two related types of models of the time series of prices. The first is a continuous-time representation; price changes can take place over any interval $[t, t + \Delta t]$, and in the absence of constraints the price change over such an interval is taken to be independent of the change in non-overlapping intervals, to have mean zero, and to have standard deviation $c\sqrt{\Delta t}$ (variance $c^2\Delta t$), where c is a "diffusion constant." Then by the Central Limit Theorem[1] the price change over any interval will be Gaussian or normal. This well-known model has been extensively studied under the popular name of "Brownian motion" and the technical name of "Wiener process."

CALCUTTA UNIVERSITY
CENTRAL LIBRARY

To standardize our analysis, we assume that units and scales are chosen so that initially $t = 0$; the fixed future time (FFT) is $t = 1$; $c = 1$; and prices are measured relative to the initial price, so that the standardized initial price is zero. Thus in the absence of additional information or constraints on the FFT price, the price at the FFT will be normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$. If ω is a path, or sequence of prices, then the price at time t will be denoted $W(t,\omega)$.

To model our two bond-selling problems, we want to constrain the Wiener process so that it has the specified distribution at the FFT; that is, so that $W(1,\omega)$ will have the desired distribution. The mathematical technique by which this may be done is rather intricate, but an intuitive explanation is as follows: let $N(x)$ be the standard normal distribution ($\mu = 0$, $\sigma = 1$) and $P(x)$ the specified distribution of FFT price. Let Ω_x be the set of paths ω for which the final price $W(1,\omega)$ is in the infinitesimal interval $[x, x + dx]$. Then for the unconstrained Wiener process the probability that a path will be in Ω_x is $N(x)\cdot dx$. To obtain the constrained process, rescale the probabilities of all the paths in Ω_x by the factor $P^*(x) = P(x)/N(x)$. The paths in Ω_x will retain their relative probabilities of occurrence, but the probability that a path of the constrained process will lie in Ω_x is now $P(x) \cdot dx$, and the overall FFT distribution is P. With this procedure we are able to constrain the process to a desired distribution at a future time and retain a random-walk type of behavior in the short run.

---

[1] See [5], pp. 229, 239.

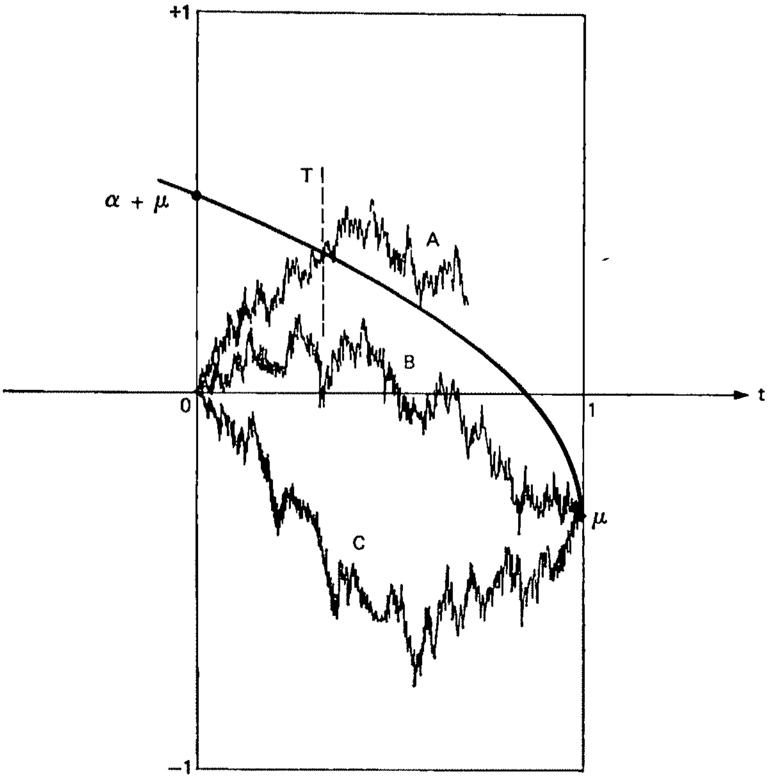
Although many attributes of the Wiener process are well-understood, our problem of finding a "stopping rule" for maximizing expected value at the stopping (or selling) time has proved to be very difficult. The results that are mathematically conclusive are due to Shepp and are stated in Theorems 1–3.

THEOREM 1: (L. A. Shepp) *If the value of the Wiener process at* $t = 1$ *is constrained to be* μ *(i.e., distributed with mean* μ *and standard deviation* $\sigma = 0$*), then the optimal strategy is: stop the first time T the value of the process is greater than or equal to* $\alpha\sqrt{1 - T} + \mu$*, where* $\alpha \approx 0.84$.

A proof of this theorem, along with the precise definition of α as the solution to an integral equation, is given in [8].

A graphical portrayal of the strategy given by Shepp appears in Figure 1. The parabola is the graph of $\alpha\sqrt{1 - t} + \mu$. On path A you pass through the stopping curve at time T and stop (sell), getting payoff $\alpha\sqrt{1 - T} + \mu$. On path B you initially have a gain, but, hoping for a larger gain, you continue and eventually have to settle

FIGURE 1



for about μ. On C you dip below μ, but you keep playing, confident that you will get at least μ. The combination of some large gains with many small losses gives you a positive expected value provided that $\alpha + \mu > 0$, i.e., if the starting point is below the stopping curve. Shepp's theorem confirms the idea that the short-term fluctuations can be exploited for a gain even when the trend is disadvantageous. We note that this game, with zero variation at the fixed future time, is analogous to the first bond seller's problem. The strategy suggested is that of "sell on a strong rally but ride out the storms," where the requirement for a rally to be considered a "strong rally" weakens as the FFT gets closer.

Earlier we mentioned that our models may often be represented as a random walk plus a trend. The following unpublished theorem of Shepp describes these cases and is needed for Theorem 3 as well as later on. The proof is rather technical and is omitted.

THEOREM 2: (L. A. Shepp) *If the value of the Wiener process at* $t = 1$ *is constrained to be Gaussian with mean* μ *and variance* $\sigma^2 \geq 1$, *then the resulting process may be represented as* $W(t,\omega) + D \cdot t$, *where* D *has a Gaussian distribution with mean* μ *and variance* $\sigma^2 - 1$.

Note that of the total FFT variance of σ^2, the Wiener process contributes 1 and the random drift D contributes $\sigma^2 - 1$. Using Theorem 2 we can obtain the stopping rule of the next theorem.

THEOREM 3: *If the distribution of the value of the Wiener process at* $t = 1$ *is constrained to be Gaussian with mean* μ *and standard deviation* $\sigma = 1$, *then for* $\mu = 0$ *any strategy gives an expected value of zero, while the optimal strategy for* $\mu \neq 0$ *is*:

(a) *stop immediately if* $\mu < 0$ *(value of game zero);*
(*b*) *continue to* $t = 1$ *if* $\mu > 0$ *(value of game* μ*).*

PROOF: By Theorem 2 the process may be represented as $W(t,\omega) + D \cdot t$, where D has mean μ and variance $1 - 1 = 0$. Since D has variance 0, $D = \mu$ and the process is $W(t,\omega) + \mu \cdot t$. If $T = T(\omega)$ is a stopping rule, then the value of the game is $E(W(T,\omega) + \mu \cdot T) = E(W(T,\omega)) + \mu \cdot E(T)$. But from [7] we have that $E(W(T,\omega)) = 0$ for any $T \leq 1$, so the maximizing T is $T = 1$ when $\mu > 0$, $T = 0$ when $\mu < 0$, and arbitrary when $\mu = 0$.

We note that although the problem in Theorem 3 is somewhat analogous to the second bond seller's problem, it is a highly restricted case, and the optimal strategy is not nearly as interesting or enlightening as that given in Theorem 1.

Theorems 1 and 3 are the only analytical solutions available to the stopping-rule problem for the Wiener process with a constraint at $t = 1$, so if we want answers when the FFT distribution is other than Gaussian with $\sigma = 1$ or a point ($\sigma = 0$), we will have to solve the problem numerically. One numerical approach would be to formulate strategies and compare them using Monte-Carlo generation of paths, taking the best found as "optimal." This method would not only be time-consuming but would limit the possible strategies to those that could be imagined by the investigators. Fortunately, with the finite or "urn" models to be discussed next we may obtain not only an efficient computational procedure but the chance of being surprised by the optimal strategy produced. These urn models are good numerical approximations to the Wiener-process model, but in their own right they form the second type of model for the original time series of prices. Thus, they are useful both for analyzing mathematical Wiener-process stopping-rule problems and for direct application to situations in economics and finance.

4. The finite or "urn" model

■ In this second method, the problem is modeled as a discrete or finite game—one with a finite maximum number of time steps, a finite number of possible price changes at each step, and a finite number of possible decisions at each step, so that there are only a finite number of possible pure strategies. A well-known approximation of this type for the unconstrained Wiener process is the tossing of an unbiased or "fair" coin.[2] If we count a head as $+1$ and a tail as -1, and keep track of the cumulative score s_k by number of tosses k, then the sequences of points (k,s_k) for $k \leq n$ will approximate Wiener-process paths on the interval $[0,n]$.[3]

The same approximation can be obtained by imagining a jar or urn filled with n balls,[4] each of which has a value of $+1$ or -1 with probability $\frac{1}{2}$. (We call those of value $+1$ "plus" balls, the rest "minus" balls.) If a player draws randomly from the urn without replacement, and we keep track of the cumulative score s_k by number of draws k, we get the same sequences and probabilities as in the coin tossing problem.

In either case the final score s_n can take on alternate integer values between $-n$ and $+n$, thus $n+1$ possible values, and the distribution of s_n is symmetric and binomial. The probability of m -1's occurring, and hence a score of $(n-m)-m = n-2m$, is given by

$$\binom{n}{m} \cdot 2^{-n} = \frac{n!}{(n-m)!m!} \cdot 2^{-n},$$

the symbol $\binom{n}{m}$ denoting a binomial coefficient, and the final score has mean zero and standard deviation $\sqrt{n}$.

If for the urn case we assume that the value of each ball was fixed before the drawing began but unknown to the player, then a final score of $n - 2m$ must have resulted from the original composition being m minus balls and $n - m$ plus balls. Thus the distribution of original compositions can be derived from the distribution of final scores and vice versa. In the urn described above, the distribution of the number m of minus balls has mean $\mu = n/2$ and standard deviation $\sigma = \sqrt{n}/2$.

Shepp used the correspondence between the original composition of the urn and the distribution of scores at the FFT (after n draws) to pose a finite stopping-rule problem analogous to the continuous-time problem analyzed in Theorem 1. There the Wiener process was constrained or "pinned" at $t = 1$ to the single possible value μ, so for the analogous finite problem there is exactly one possible final score, $n - 2m$. But this implies that there were exactly m minus balls and $n - m = p$ plus balls in the original urn. Thus, "Shepp's urn problem" [8] is to define a stopping rule for randomly drawing without replacement from an urn of known composition so as to maximize the expected score at the stopping time. This is a finite analogy of the first bond seller's problem.

---

[2] See [5], ch. III.
[3] *Ibid.*, p. 84.
[4] *Ibid.*, p. 108 ff.

The author proposes to generalize Shepp's problem to allow an arbitrary probability distribution on the $n + 1$ possible final scores. Since a distribution on the final scores implies a distribution on the original composition of the urn, the random urn problem is to find an optimal stopping rule, given the total number of balls and a probability distribution of the number of minus (or plus) balls in the urn. This problem is then a finite analogy of the second bond seller's problem.

Shepp's urn problem is discussed in section 5 and the random urn problem in section 6. Details of the analysis will be found in Appendix 1 or in [2].

5. Shepp's urn problem

■ We shall denote by $V(m,p)$ the value under optimal play (optimizing expected value) of an urn with n balls, m of them minus (value -1) and $n - m = p$ of them plus (value $+1$). The rules of the game are that the player is told the composition of the urn; then he can draw from zero to n of the balls, randomly, one at a time, without replacement. Thus he begins with all n in the urn and can stop any time, including before drawing even one ball. When he stops he gets a score of $+1$ for each plus ball drawn and -1 for each minus ball. The player's objective is to draw in such a way as to maximize his expected score upon stopping. We have of course $V(m,p) \leq p$, the number of plus balls, and since the player need never draw he will not accept a negative expectation, so $V(m,p) \geq 0$. Thus $V(m,0) = 0$, and taking all balls gives $V(0,p) = p$.

Shepp has pointed out[5] that $V(m,p)$ satisfies a relatively simple forward recursion formula (Theorem 5 of Appendix 1). Thus if we know the value $V(m,p)$ for all cases with $m + p = n$, we can compute the value for all urns with $n + 1$ balls using backward induction, or

TABLE 1.
V (m,p)

p (PLUS)										
9	9.	8 10	7.20	6.31	5 43	4 58	3.75	2.95	2.21	1.53
8	8.	7 11	6.22	5.35	4 49	3.66	2 86	2.11	1.43	0 84
7	7.	6.13	5.25	4 39	3 56	2 76	2.01	1 34	0.75	0.30
6	6	5 14	4.29	3 45	2 66	1.91	1 23	0.66	0 23	0
5	5	4.17	3.33	2 54	1 79	1.12	0.55	0.15	0	0
4	4	3 20	2.40	1 66	1 00	0.44	0.07	0	0	0
3	3	2.25	1.50	0 85	0.34	0	0	0	0	0
2	2.	1.33	0 67	0.20	0	0	0	0	0	0
1	1	0 50	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
	0	1	2	3	4	5	6	7	8	9
	m (MINUS)									

---

[5] See [8], p. 1000.

"backtrack," since we can easily calculate the probabilities of the first draw. The details are given in Appendix 1.

Numerical results for the value of the (m,p) urns for small values of m and p are given in Table 1. Fixed values of $n = m + p$ occur as top-left to lower-right diagonals. The entries for $m = p$ are underlined.

We observe from Table 1 that there are many urns with $m > p$ which have a positive value. In fact, as a corollary to Theorem 1 Shepp was able to show that if $\beta(p)$ is the largest m for which $V(m,p) > 0$, then for large p,

$$\beta(p) \approx p + \alpha(2p)^{\frac{1}{2}},$$

where $\alpha \approx 0.84$ as in Theorem 1. Thus in the finite case also the short-term fluctuations can be exploited for a gain even when the overall trend is unfavorable.

We can use Table 1 to play an (m,p) urn optimally; we begin by placing a marker on the (m,p) square. If the value there is zero, an optimal strategy is to stop immediately, that is, drawing a ball has an expected payoff which is zero or negative even when the value of the chance to keep drawing is included. If the value of the urn is positive, we should draw. If a plus ball is drawn, the reduced or "new" urn is the $(m, p - 1)$ urn, so we should move the marker one square down and proceed as before; if we draw a minus ball, we should move the marker one square to the left. As soon as we come to a "zero square," we should quit.

Thus all we really need to play optimally is a "stopping set," the set of urns with value zero. The exact values of the positive-valued urns are only intermediate results. We observe that an initial string of successes will cause us to quit early, while with initial losses we tend to keep playing. Thus the overall strategy is similar to that of Theorem 1 (as it should be), namely, sell on strong rallies but ride out the storms.

6. The random urn problem

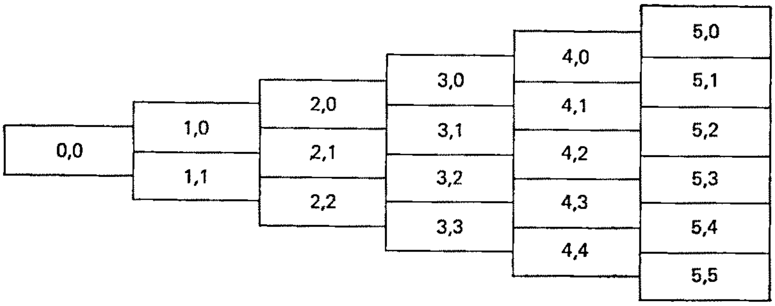
■ Suppose now that instead of an (m, p) urn we are given an $(n; P)$ urn, when n is the total number of balls in the urn, which is still assumed to be known, and P is a probability distribution on the $n + 1$ integers $0, 1, 2, \ldots, n$, with $P(m)$ being the probability that m of the balls are minus (have value -1) and $n - m$ are plus (value $+1$). We noted earlier that P also gives the distribution of the final score if all n balls are drawn, with $P(m)$ being the probability that the final score would be $n - 2m$. The problem is, as before, to define an optimal stopping rule or strategy for maximizing the expected score at the stopping time. We denote the value of the $(n; P)$ urn under optimal play by $V(n; P)$.

In principle the backtrack method can be applied to this problem as it was to the known-urn problem. The drawing of a ball from an urn with $n + 1$ balls still reduces the urn to one with n balls, and the chance of drawing a plus or minus ball can be calculated easily. But since there are an infinite number of random urns with n balls, in contrast to only $n + 1$ possible known-composition urns, we can't expect to have the necessary values of the reduced n-urns already computed. Thus no matter how many urn problems have been solved previously, to solve a new random-urn problem by Shepp's backtrack method it is virtually certain that we will have to start from scratch and solve a cascade of reduced-urn problems which are of little or no interest in themselves.

We conclude that a more appropriate approach to solving the random urn problem is to seek an algorithm for producing an *ad hoc*, explicit strategy for the random urn of interest, rather than using a strategy described in terms of values or strategies for all possible reduced urns. In the rest of this section we present such an algorithm.

To describe our strategy we shall use a triangular table, such as that shown for a 5-urn in Figure 2. We compute two quantities a and b for each square of the table. Each of the squares is identified by a pair of integers (i,j), with $0 \leq i \leq n$ and $0 \leq j \leq i$. The quantities $a(i,j)$ and $b(i,j)$ will describe the game situation after i balls have been drawn, of which j were minus balls and $i-j$ plus balls. The (0,0) square thus represents the beginning of the game, and the $n+1$ squares in the last column represent the possible final outcomes of the game if all n balls are drawn. To apply the optimal strategy, we place a marker on the (0,0) square, then move the marker right and down when a minus ball is drawn, and right and up when a plus ball is drawn. The decision whether to draw again after drawing i balls, j of them minus (that is, after reaching the (i,j) square), depends on whether $b(i,j)$ is positive or zero, just as we use $V(m,p)$ from Table 1 for Shepp's urn problem. But in this case the table used is made up especially for the problem at hand, while Table 1 (or its extension to larger m and p) suffices for all possible urns of known composition.

FIGURE 2



The algorithm for computing $b(i,j)$, and thus the optimal strategy and value $V(n;P)$ for an $(n;P)$ urn, is given in Theorem 4. The derivation of this algorithm is given in Appendix 2. There we see that the procedure works because $b(i,j)$ is a scaled form of the remaining value at the (i,j) point in the game. The quantity $a(i,j)$ is a similarly scaled form of the expected value of the $(i+1)^{st}$ draw and is used in the computation of $b(i,j)$.

THEOREM 4: *Given an $(n;P)$ urn, for $0 \leq j \leq n$ let $P^*(j) = P(j)/\binom{n}{j}$ and $a(n,j) = b(n,j) = 0$. For $0 \leq j \leq n-1$ let $a(n-1,j) = P^*(j) - P^*(j+1)$, and for $0 \leq i \leq n-2$ and each j let*

$$a(i,j) = a(i+1,j) + a(i+1,j+1).$$

For $0 \leq i \leq n-1$ and each j let

$$b(i,j) = \max\{0, a(i,j) + b(i+1,j) + b(i+1,j+1)\}.$$

Then an optimal stopping rule for maximizing the expected score at the stopping time is: if i balls have been drawn, j of them minus, draw again only if b(i,j) > 0. The value V(n;P) of the (n;P) urn under optimal play is b(0,0).

For a numerical example of the use of Theorem 4, let us take $n = 5$ and let P be the equiprobable distribution:

$$P(j) = 1/6, \quad 0 \le j \le 5 .$$

Then we have

$$P^*(0) = P^*(5) = 1/6 = 10/60 \ ,$$
$$P^*(1) = P^*(4) = 1/30 = 2/60 ,$$
$$P^*(2) = P^*(3) = 1/60 ,$$

and the quantities $a(i,j)$ and $b(i,j)$ are tabulated in Figures 3 and 4, in the format of Figure 2. In Figure 3 we also give columns for j and $P^*(j)$. We see that $V(5;P) = b(0,0) = 42/60$.

FIGURE 3

a(i,j) AND P*(j)

						j	P*(j)
0	10/60	10/60	9/60	8/60	0	0	10/60
	−10/60	0	1/60	1/60	0	1	2/60
		−10/60	−1/60	0	0	2	1/60
			−9/60	−1/60	0	3	1/60
				−8/60	0	4	2/60
					0	5	10/60

FIGURE 4

b (i,j)

42/60	42/60	30/60	18/60	8/60	0*
	0*	2/60	2/60	1/60	0*
		0	0*	0*	0*
			0	0	0
				0	0
					0

The last column in each triangle is zero by definition. The next-to-the-last column in the $a(i,j)$ triangle is the difference of the nearest values of $P^*(j)$. The $a(i,j)$'s add pairwise from right to left to fill in the triangle. An entry in the $b(i,j)$ triangle, say $b(2,0) = 30/60$, is computed from

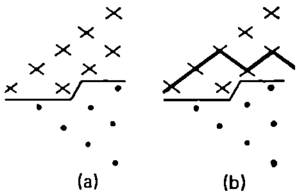
$$b(2,0) = \max\{0, 10/60 + 18/60 + 2/60\} ,$$

where $a(2,0) = 10/60$ and $18/60$ and $2/60$ are the values of $b(i,j)$ adjacent at the right. If the sum were negative instead of $30/60$ (due to $a(2,0)$ being negative), we would set $b(2,0) = 0$.

We see that the only squares at which we will stop while playing optimally are those marked with a star (*). The optimal strategy consists of stopping if the first ball is minus, otherwise stopping whenever two minus balls have been drawn or the urn is empty. Thus, unless all balls are drawn, the last ball drawn will be a minus. This contrasts with Shepp's urn problem, for which the last ball drawn under optimal play is always a plus (see Appendix 1). As might be expected, the random urn strategies show considerably more variation than the strategies for Shepp's urn problem.

For computer plots in the next section we abstract strategy information such as in Figure 4 by deleting the last column, then letting (X) denote a square with $b(i,j) > 0$ and ($\cdot$) a square with $b(i,j) = 0$, as shown in Figure 5a. In these plots a square marked (0) cannot be reached; for instance, if there were at most three minus balls in a 5-urn, we will plot a (0) for square (4,4). The light line in Figure 5 denotes the stopping boundary. The heavy line in Figure 5b illustrates the path resulting from drawing $+1, +1, -1, +1, -1$.

FIGURE 5


(a) (b)

7. Numerical investigations

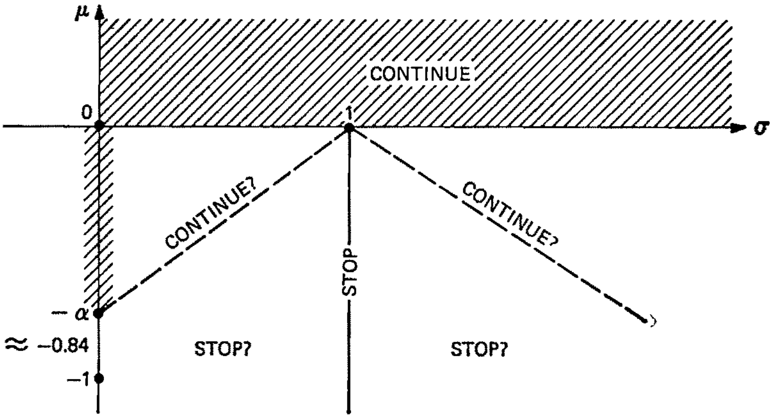
■ The algorithm of Theorem 4 has been implemented as a simple and effective computer program which computes values and optimal strategies for a wide range of $(n; P)$ urns. Most of the cases we have studied use a P which is discretized Gaussian; the program accepts a desired μ and σ and computes the corresponding P. An arbitrary table for P may also be entered, and the program has been modified from time to time to generate special distributions given control parameters.

Our principal objective has been to obtain generalizations of Shepp's theorem on the "pinned" Wiener process. Since his Theorem 1 solved a problem analogous to the first bond seller's problem, we believed that more widely applicable theorems would shed some light on the second bond seller's problem, where only a probability distribution is known for the price at the fixed future time (FFT).

Almost all our work so far has been on the case in which the FFT price distribution is Gaussian; Theorem 2 indicates that this is an important case. A variety of problems can be defined by varying the mean μ and standard deviation σ of the FFT distribution. Both Theorems 1 and 3 fit nicely into this program, Theorem 1 disposing of all μ for $\sigma = 0$ and Theorem 3 covering all μ for $\sigma = 1$.

Our first goal was to determine "which urns are favorable"; that is, for what values of σ and μ does a Gaussian distribution at the FFT provide a positive expected value under optimal stopping. We shall display our conclusions as shown in Figure 6, which illustrates the conclusions of Theorems 1 and 3. The shaded area indicates those points (σ,μ) for which we know the value $V(\sigma,\mu)$ is positive. The point $(0, -\alpha)$ divides $V(0,\mu) > 0$ from $V(0,\mu) = 0$ by Theorem 1; similarly, from Theorem 3 we know that the point (1,0) divides $V(1,\mu) > 0$ from $V(1,\mu) = 0$. The first quadrant is shaded since the opportunity of playing to the end ensures that $V(\sigma,\mu) \geq \mu$, so $V(\sigma,\mu) > 0$ when $\mu > 0$. The broken line plots a hypothetical minimum μ as a function of σ, $\mu = \mu(\sigma)$; it seemed reasonable to conjecture that our isolated theoretical points $(0, -\alpha)$ and (1,0) lay on a single continuous curve. There was still the possibility that the minimum μ was zero for $\sigma > 1$, but from Theorem 2 we know that in

FIGURE 6



that case the process could be represented as a random walk plus an unknown drift or trend. Thus, it seemed plausible that by estimating the drift one could achieve $V(\sigma,\mu) > 0$ for some $\sigma > 1$ and $\mu < 0$.

So, given a pair of parameters (σ,μ), we wish to know whether $V(\sigma,\mu) > 0$. To answer this question we select n and P so that the "sign" (plus or zero) of $V(n; P)$ is a good estimate of the "sign" of $V(\sigma,\mu)$. For the Wiener-process model the distribution of values at the FFT is Gaussian; thus the distribution of values (final scores) at the FFT of the urn approximation should be close to Gaussian. But since the final scores are just $n - 2m$, where m is the number of minus balls originally in the urn (given by P), the distribution P should also be close to Gaussian, with parameters μ' and σ'. It remains to compute the parameters σ' and μ' and the distribution P from σ, μ, and n.

The key to this transformation is found in the analogy we described earlier between the unconstrained Wiener process on $[0,n]$ and drawing from an $(n; \bar{P})$ urn where $\bar{P}$ is symmetric and binomial, with each ball having the values $+1$ and -1 with probability $\frac{1}{2}$. We will pick σ' and μ' so that P bears the same relationship to the binomially-distributed $\bar{P}$ as the (σ,μ) distribution bears to the standard $(1,0)$ distribution N. This is accomplished through the transformations

$$\sigma' = \sigma\sqrt{n}/2$$

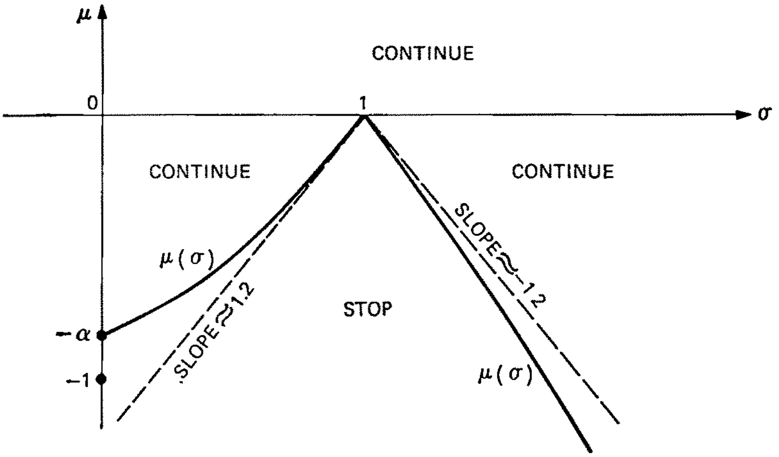
$$\mu' = -\mu\sqrt{n}/2 + n/2.$$

Then for a given n, P is obtained by discretizing the (σ',μ') Gaussian distribution.

Of course, n is chosen as a compromise between accuracy and computational feasibility. Our original program gave reasonable accuracy with $n \approx 100$; a newer, more sophisticated version can handle n around 1000, furnishing what we believe to be excellent precision. Computation of $V(n; P)$ using Theorem 4 takes about $8n^2$ microseconds on a CDC 6600.

Figure 7 presents our computation of the minimum μ as a function of σ, $\mu = \mu(\sigma)$. For all points (σ,μ) above the graph, $V(\sigma,\mu) > 0$, while on the graph and below, $V(\sigma,\mu) = 0$. The numbers plotted are given in Table 2. The large areas for which the value is positive although $\mu < 0$ make it clear that for this class of problems also, it is often possible to exploit the variation in price for a profit when the trend is unfavorable.

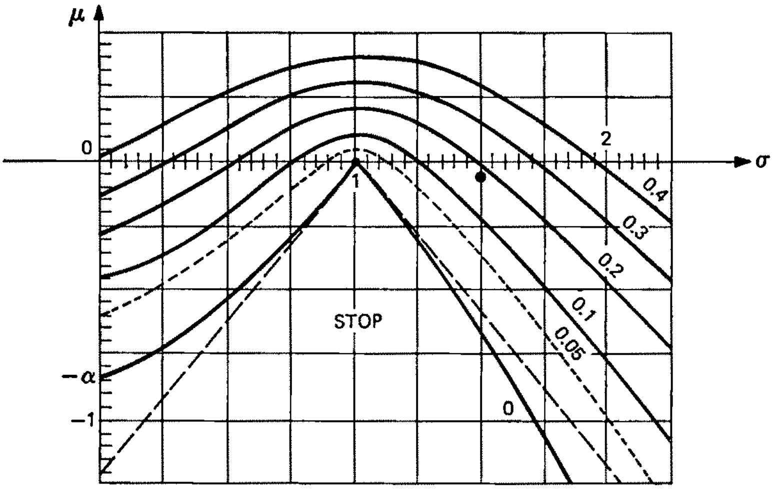
FIGURE 7



Two aspects of Figure 7 deserve comment. The approximations are quite poor near $\sigma = 0$, but we are convinced that $\mu(\sigma)$ is continuous at $\sigma = 0$; however, we are unable to estimate the slope there. The slopes of the left and right tangents at $\sigma = 1$ appear to have the same absolute value, about 1.2. A very likely value for the slope is 1.28 $\approx 2 \cdot \gamma_1$, where γ_1 is the solution of an integral equation. The constant γ_1 will arise again in a similar context in the next section.

For later applications it will be useful to have an estimate of the value $V(\sigma,\mu)$ at those points for which it is positive. Contour lines for $V(\sigma,\mu)$ are given in Figure 8.

FIGURE 8



Once we had an idea of which (σ, μ) urns were favorable, our next question was: "How does the optimal strategy depend on σ and μ?" This led to what we believe is our most interesting conclusion. Perhaps we should have suspected from the kink in $\mu(\sigma)$ at $\sigma = 1$ that typical strategies for $\sigma < 1$ and $\sigma > 1$ would be different. Indeed, Figures 9–11 reveal that this is the case. They are plotted as described at the end of the last section, with (X) marking a playing square ($b(i,j) > 0$), ($\cdot$) a stopping square, (0) denoting a square which is "impossible" (cannot occur for the given P), and $n = 64$.

Figure 9 shows $(\sigma,\mu) = (0, -0.5)$ a "known-composition" case; as we saw in Theorem 1, the stopping boundary is an upper parabola, and the strategy is "sell on strong rallies, ride out storms." Figure 10 shows $(\sigma,\mu) = (0.5, -0.5)$, and the strategy is similar. But in Figure 11, with $(\sigma,\mu) = (1.5, -0.5)$, the strategy has "flipped over!" The stopping boundary is now *below*; the strategy has become "cut losses, let profits run!"

FIGURE 9

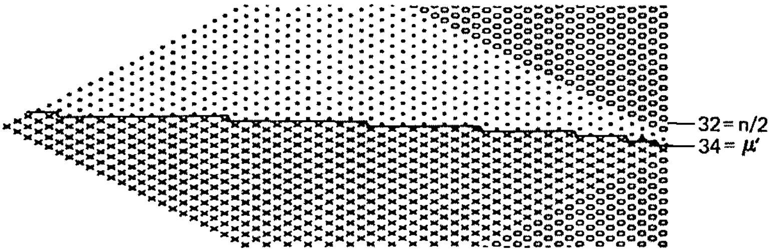


FIGURE 10

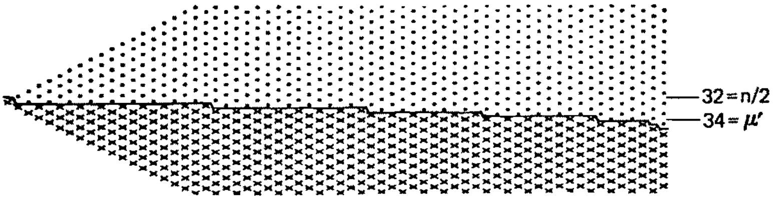
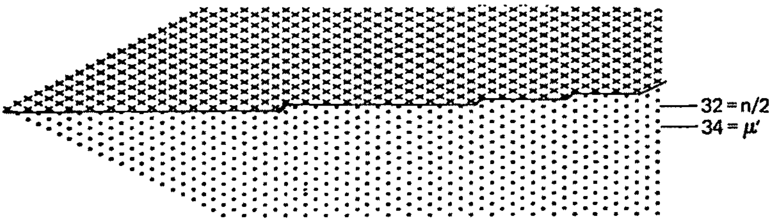


FIGURE 11



The contrast between typical $\sigma > 1$ and $\sigma < 1$ strategies is particularly interesting in that the "let profits run" motto seems to be considerably more successful and widely quoted in the financial world than "ride out storms." This situation leads one to believe that the $\sigma > 1$ model is more appropriate to the real world; the appealing interpretation provided for this case by Shepp in Theorem 2 lends credence to this view.

8. Shape of the stopping boundary

■ It is difficult to estimate the functional form of the stopping boundary in Figures 9–11; it could conceivably be a straight line, parabola, or a more esoteric curve. The lack of distinctness of the stopping boundary is due to the nature of the urn approximation to the Wiener process. The horizontal coordinate (time) is scaled by $1/n$, but the vertical coordinate (score or price) is scaled by $1/\sqrt{n}$. Thus the vertical approximation in Figures 9–11 is $\sqrt{64} = 8$ times rougher than the horizontal, and only 16 different vertical values are available in the important $[-1, 1]$ interval. The idea of developing the precise shape of the stopping boundary from plots such as these appears to be naive.

For the Gaussian case, however, L. A. Shepp, C. L. Mallows, and the author have recently derived theoretical transformations which allow the stopping boundary to be computed with considerable accuracy. We assume (on convincing evidence) that at each time t there is exactly one boundary point; that is, the boundaries in Figures 9–11 are typical Gaussian boundaries. Then for each pair of parameters σ and μ, and for each t, $0 \leq t \leq 1$, we indicate the boundary point (t,x) by $x = \beta(\sigma,\mu,t)$ and call β the boundary function.

By considering certain properties of the Wiener process with a Gaussian constraint at $t = 1$, we conclude that β must satisfy the following identity:

$$\beta(\sigma,\mu,t) = \mu/(1 - \sigma^2) + (1 - t)^{\frac{1}{2}} \cdot \beta(\sigma(1 - t + t\sigma^2)^{-\frac{1}{2}}, 0, 0)\,.$$

Thus to compute the boundary it suffices to be able to evaluate the boundary function when $t = 0$ and $\mu = 0$. We denote this restricted function by $\gamma(\sigma)$:

$$\gamma(\sigma) = \beta(\sigma,0,0)\,.$$

If $\mu = \mu(\sigma)$, the lower bound of the μ for which $V(\sigma,\mu) > 0$, then the initial point $t = 0$, $x = 0$ must lie on the stopping boundary; so with $t = 0$, $\mu = \mu(\sigma)$ we have $x = 0$, $(1 - t)^{\frac{1}{2}} = 1$, and $\sigma(1 - t + t\sigma^2)^{-\frac{1}{2}} = \sigma$, so

$$0 = \beta(\sigma, \mu(\sigma), 0) = \mu(\sigma)/(1 - \sigma^2) + \gamma(\sigma)\,.$$

Thus we may compute $\gamma(\sigma)$ from

$$\gamma(\sigma) = -\mu(\sigma)/(1 - \sigma^2)\,.$$

Table 2 gives the values of $\mu(\sigma)$ obtained numerically and the resulting values of $\gamma(\sigma)$.

It is curious that although $\gamma(\sigma)$ is discontinuous at $\sigma = 1$, $|\gamma(\sigma)|$ appears to be continuous and smooth. The interpolated value $|\gamma(1)| = 0.623$ is very close to the theoretical value $\gamma_1 \approx 0.639$ obtained by B. F. Logan (unpublished); γ_1 is the solution of an integral equation similar to the equation which defines α.

The character of the stopping boundary is seen most easily by examining the family of boundaries for $\mu = 0$, which all go through $x = 0$ at $t = 1$. They are graphed in Figure 12. If $|\gamma(\sigma)|$ is monotone, as it must be, the boundaries move closer to the t-axis as σ increases, and they seem to intersect only at $t = 1$. Shepp's parabola $\alpha\sqrt{1 - t}$ is an upper bound for the boundaries with $0 \leq \sigma < 1$. The upper branch of the parabola $\gamma_1\sqrt{1 - t}$ is a lower bound for $0 \leq \sigma < 1$, while the lower branch of the same parabola is the lower bound for

TABLE 2.
$\mu(\sigma)$ AND $\gamma(\sigma)$

σ	n	$\mu(\sigma)$	$\gamma(\sigma)$
0	–	–0.84*	0.84*
0.10	625	–0.802	0.804
0.30	625	–0.693	0.762
0.50	625	–0.538	0.717
0.70	625	–0.344	0.675
0.90	625	–0.121	0.638
0.95	1024	–0.062	0.633
1.00	–	0*	±0.623** ±0.639*
1.05	625	–0.063	–0.613
1.10	625	–0.127	–0.605
1.30	625	–0.396	–0.574
1.50	1024	–0.689	–0.551
2.00	1024	–1.485	–0.495
2.50	1024	–2.372	–0.452
3.00	1024	–3.348	–0.419
4.00	1024	–5.606	–0.374
5.00	1024	–8.336	–0.347
6.00	1024	–11.592	–0.331
7.00	1024	–15.402	–0.321

*THEORETICAL
**INTERPOLATED

$1 < \sigma < \infty$. The boundaries cease being monotone functions of t for σ beyond about 2. The upper bound for the $\sigma > 1$ cases, which might be called the "$\sigma = \infty$ boundary," is interesting. As $\sigma \to \infty$,

$$\sigma(1 - t + t\sigma^2)^{-\frac{1}{2}} = \left(\frac{1-t}{\sigma^2} + t\right)^{-\frac{1}{2}} \to t^{-\frac{1}{2}},$$

FIGURE 12

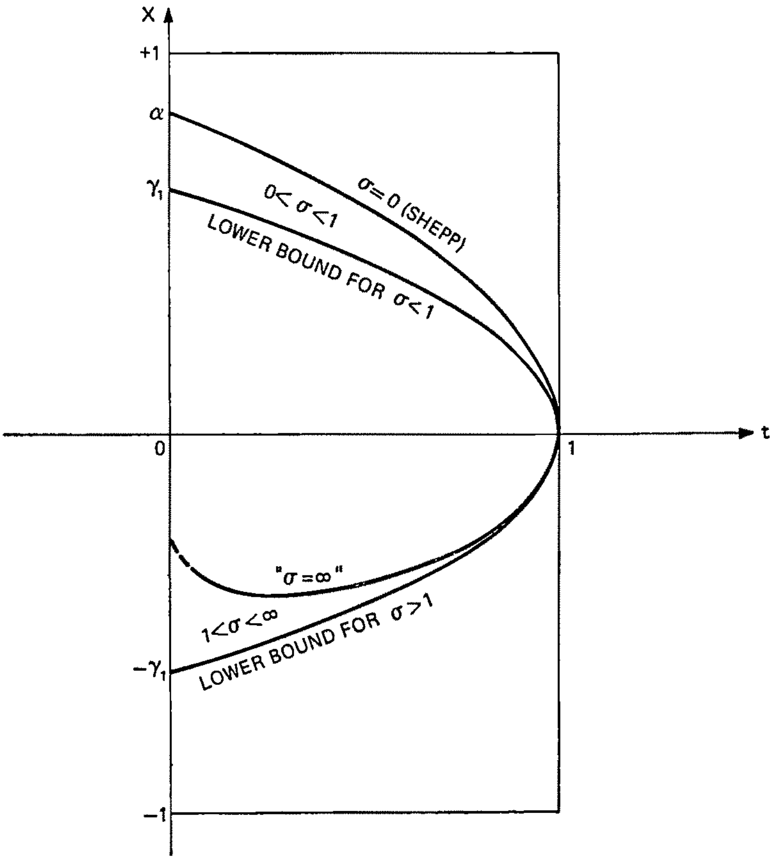
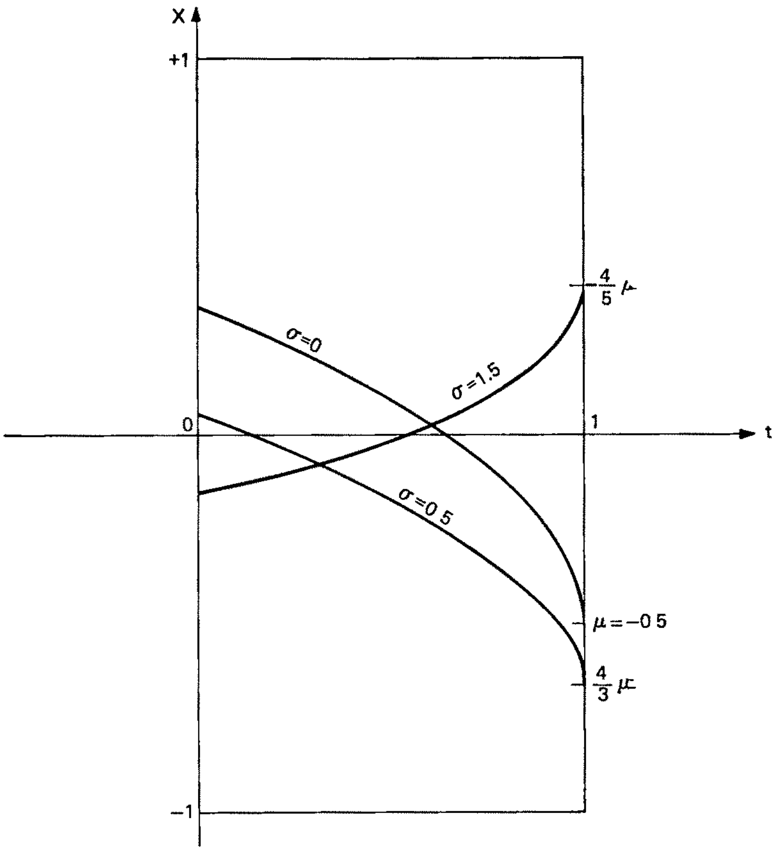


FIGURE 13



so

$$\beta(\infty,0,t) = (1-t)^{\frac{1}{2}}\gamma(t^{-\frac{1}{2}}).$$

As t goes from 0 to 1, $t^{-\frac{1}{2}}$ goes from $+\infty$ to 1, so we can estimate the $\sigma = \infty$ boundary closely when $t^{-\frac{1}{2}} \leq 7$, or $t \geq 0.02$.

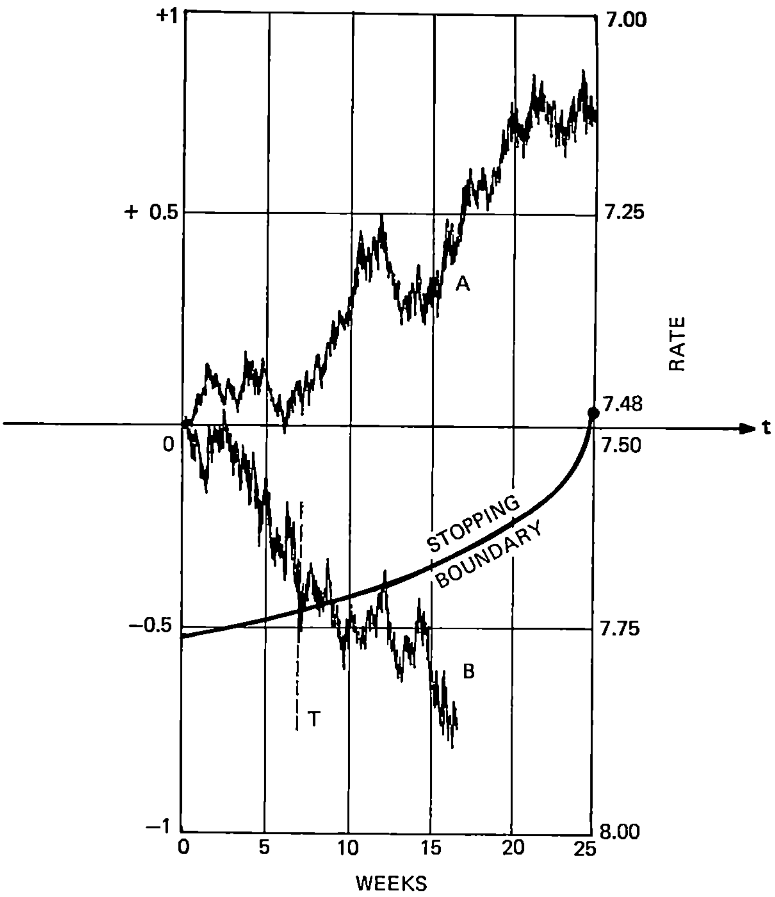
The boundaries for $\mu \neq 0$ are obtained by rigidly shifting the $\mu = 0$ boundary up or down until it hits the FFT at $\mu/(1 - \sigma^2)$. The resulting family of boundaries for fixed μ and varying σ can be a confusing jumble when $\mu < 0$. Figure 13 shows three boundaries for $\mu = -0.5$; they are the continuous-time versions of the boundaries in Figures 9–11.

9. An example

■ As an illustrative exercise in scaling coordinates and using the graphs and formulas presented earlier, we work out a solution for a commonplace financial problem. We wish to *minimize* the (expected) interest rate on a bond issue. In our previous discussions we wanted to *maximize* bond prices, which is equivalent, and the common practice of plotting yields on a reversed scale points up the equivalence. However, at one point we will have to change a sign in order to read a graph properly, and we hope this will not unduly disturb the reader.

We assume a current interest rate of 7.50 percent (the exact figure is immaterial) and that the weekly change in interest rate has a standard deviation of 0.1 percent. Thus the weekly variance is 0.01. The bonds must be issued (or bids received) within 25 weeks. We

FIGURE 14



believe that over the 25-week period there will be a drift or trend in interest rates which is unknown, but which measured weekly is Gaussian with mean of +0.001 percent and standard deviation 0.022 percent.

The variance of the unconstrained (driftless) process over 25 weeks is $25 \times 0.01 = 0.25$, so the standard deviation is 0.50%. Thus, our vertical unit of scale will be 0.50%. Our horizontal unit of scale is 25 weeks, so one week is $t = 0.04$. Over the 25-week period the drift will have mean $25 \times 0.001 = 0.025\%$ and standard deviation $25 \times 0.022\% = 0.55\%$. Note that for the drift, we add standard deviations instead of variances, since the same drift is assumed to be present for all 25 weeks, perfectly correlated week-to-week. The total variance at 25 weeks is $(0.50)^2 + (0.55)^2 \approx (0.75)^2$, so the total standard deviation is 0.75%. In terms of vertical units, $\sigma = 0.75/0.5 = 1.5$ and $\mu = 0.025/0.5 = 0.05$, which we interpret as -0.05 since increasing yields give decreasing prices. Now $\mu/(1 - \sigma^2) = -0.05/(-1.25) = 0.04$, which scales to -0.02%, or $7.50\% - 0.02\% = 7.48\%$ in absolute terms. The stopping boundary is shown in Figure 14. The expected gain from playing optimally may be read off Figure 8 (at the dot): $V(1.5, -0.05) \approx 0.20$ vertical units, or $0.2 \times 0.5\% = 0.10\%$, or 10 basis points. The source of the expected gain is, of course, the good chance of issuing the bonds in week 25 at a rate considerably below 7.48 percent (path A). This is not balanced by high rates, since we avoid entering situations where rates are above 7.75 percent (path B).

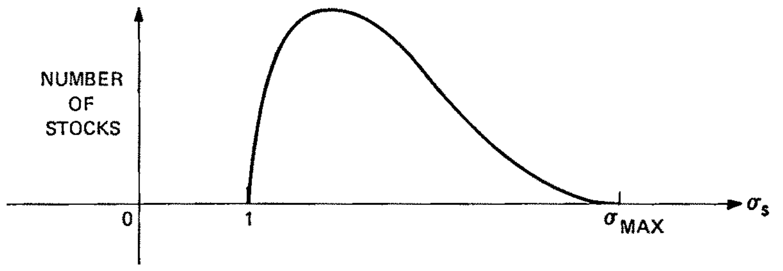
10. Stopping rules and "speculative splurges"

■ Although this paper discusses the models in terms of timing the sale of bonds, there are obviously applications to other speculative markets as well, including the major stock exchanges. In these markets a future price is seldom known exactly, so a "problem 2" model is most appropriate. In this section we discuss one particular stock-market phenomenon in terms of the Wiener process with a Gaussian FFT price distribution.

We have observed that in recent years the ending of a market upswing (bull market) often coincides with a period of intense speculative activity. The general market averages are heavily weighted with high-quality (consistent earning power, or "blue-chip") stocks, while the speculative activity is centered on low-quality stocks, stocks of companies with erratic histories or uncertain prospects. Thus the "speculative splurge" is often considered to be irrational, since blue-chip prices are generally indicative of future prospects for the economy as a whole; the speculative-stock companies must operate in this same economy, but are supposedly less capable than the blue-chip companies of conducting successful operations when economic conditions are poor. Thus it appears that the investors who bid up the speculative stocks after the blue-chip prices have stopped increasing are ignoring economic realities, and therefore it is often alleged that the speculators are "the public" or some other poorly-informed segment of the market. However, close study reveals that many well-informed professionals are speculating, too, and we will show that the Gaussian probabilistic model can be interpreted so as to indicate that the speculative stocks are the best buys when prices are starting to trend down.

To obtain this interpretation we make a number of assumptions which we specify as follows. If the present time is t_0, the change in the price of each stock S over the time interval $[t_0, t_0 + \Delta t]$ (Δt about six to nine months) will be the sum $M(S)W(t, \omega) + D(S)t$ of a multiple $M(S)$ of the Wiener process $W(t, \omega)$ and a drift $D(S)t$ with normally distributed slope $D(S)$. We scale time so that $\Delta t = 1$ and the price of each stock so that $M(S) = 1$. Then at $\Delta t = 1$ the drift will have an expected value $\mu_{D(S)}$ and standard deviation $\sigma_{D(S)}$, and the stock price will have expected value $\mu_S = \mu_{D(S)}$ and variance $\sigma_S{}^2 = 1 + \sigma_{D(S)}{}^2 \geq 1$. (We thus avoid the case $\sigma < 1$, which is more difficult

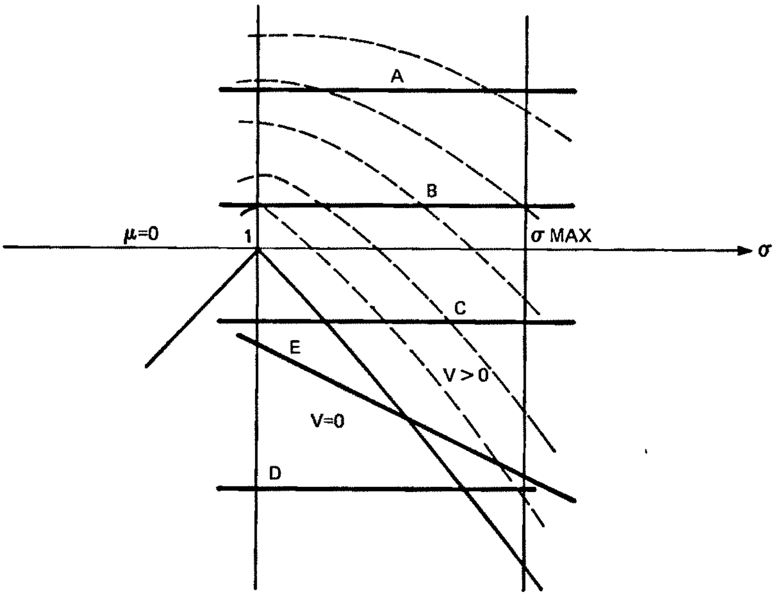
FIGURE 15



to interpret than $\sigma > 1$.) Blue-chip stocks will have more predictable performance, hence smaller $\sigma_{D(S)}$ and σ_S close to one, while speculative or unproven stocks will have larger values of σ_S. Listing regulations will keep σ_S below some $\sigma_{\max}$. Thus the distribution of stocks by σ on an exchange will be similar to Figure 15.

We restrict our attention to the interval $[1, \sigma_{\max}]$ of Figure 8, the plot of V as it depends on σ and μ. This section of the plot is magnified for Figure 16.

FIGURE 16



The lettered bars represent various levels of expected price changes over the interval $[t_0, t_0 + \Delta t]$. When economic prospects are good, with μ positive for most stocks (line A), the values of stocks with varying σ are roughly the same and prices go up together. When prospects are not so good (line B), profits can be made in stocks with all σ, but the payoff for skillful investment in high-σ (speculative) stocks is more, so speculative stocks become more favored. Finally, when overall prospects are poor (line C) or bad (line D), investments in low-σ (blue-chip) stocks produce an expected loss, while profits can still be made by skillful trading in high-σ stocks, with the quality of stocks that must be chosen declining along with the economic prospects. We note that the results still hold if one wishes to assume that lower quality (higher-σ) stocks generally have poorer expected prospects (lower-μ), as represented by line E; for E, μ is lower for the high-σ stocks, but still money can be made with them but not with the lower-σ, higher-μ blue-chips.

11. A moral on modeling

■ The building of mathematical models of economic, social, biological, and physical activities and systems is a common practice and seems destined to become even more prevalent. These models are used to provide insight into the workings of the real world, as aids for planning and operations, and occasionally as a standard by which to judge the adequacy or efficiency of real-world activities. All too often these models assume perfect knowledge when the corresponding real-world situation is rife with uncertainty. This paper presents an example of a situation in which for sufficiently great uncertainty, policies based on a model in which the future is known will be the complete opposite of the correct policies—the $\sigma = 0$ policy says "sell on rallies" while for $\sigma > 1$ we should "let profits run!" We are convinced that this conflict between deterministic and random-model policies occurs frequently, and that therefore random influences should be included in models (particularly economic models) at an early stage of development. Particularly untrustworthy, in our opinion, are those analyses which assume that the individuals involved in an economic situation have a perfect knowledge of the various parameters and functions defining the situation.

Appendix 1

Further information on Shepp's urn problem

■ For the (m, p) urn, we draw a minus ball with probability m/n, giving an immediate score change of -1 and the opportunity to draw from the $(m - 1, p)$ urn, with value $V(m - 1, p)$; while when a plus ball is drawn, with probability p/n, we get $+1$ and the reduced urn is the $(m, p - 1)$ urn. Thus the expected value $E(m,p)$ of drawing from the (m,p) urn, including the chance to make subsequent draws, is

$$E(m,p) = \frac{m}{n}(-1 + V(m - 1, p)) + \frac{p}{n}(+1 + V(m, p - 1)) .$$

Then, if $E(m,p) > 0$, we draw from the (m,p) urn, so $V(m,p) = E(m,p) > 0$; while if $E(m,p) \leq 0$, we choose not to draw, and $V(m,p) = 0$. Thus

$$V(m,p) = \max\{0, E(m,p)\}$$

and we have

THEOREM 5: (Shepp) *$V(m,p)$ is defined inductively by $V(0,p) = p$, $V(m,0) = 0$, and*

$$V(m,p) = \max\left\{0, \frac{m}{n}(-1 + V(m - 1, p)) + \frac{p}{n}(+1 + V(m, p - 1))\right\} .$$

Using this theorem we can compute the values given in Table 1. However, there is a better method, based on Theorem 4, which is given in [2].

Table 1 confirms what are intuitively obvious suppositions, that adding a good ball to the urn should increase the value of the urn and that adding a bad ball should reduce the value. These facts are surprisingly tricky to prove, but in [8] Shepp has sketched a proof;

THEOREM 6: (Shepp) $V(m,p + 1) \geq V(m,p) \geq V(m + 1,p)$.

COROLLARY: *Under optimal play, the last ball drawn is always a plus.*

PROOF: If a minus ball is drawn for the (m,p) urn, then we must have had $V(m,p) > 0$. But the reduced urn is the $(m - 1, p)$ urn, and

$V(m-1,p) \geq V(m,p) > 0$, so another ball will be drawn. Thus a minus ball is never the last drawn.

Other properties of Shepp's urn problem are discussed in [2].

Appendix 2

The proof of Theorem 4

■ In the proof we shall rely heavily on the randomness of the draws from the urn. The randomness allows us to disregard the *sequence* of draws that have gone before and keep track only of the *number* of draws and number of minus balls drawn (or equivalently, the number of draws and current score). This is the reason we can set up a table like Figure 2, which distinguishes cases only on the basis of i and j. The consequences of the randomness are stated without proof in Lemma 1.

LEMMA 1: *Let α and β be initial sequences of i draws from an $(n;P)$ urn such that in each sequence j minus balls and $i-j$ plus balls are drawn, but in possibly different orders. Then the probabilities of α and β are equal. Furthermore, if ϵ is any event which can occur on draw $i+1$ through n, the probabilities of ϵ conditioned on α and β are equal.*

Definition: Given an $(n;P)$ urn:

(1) $Q(i,j)$ is the probability that j of the first i balls drawn are minus and $i-j$ plus,

(2) $q(i,j)$ is the probability of any specific sequence α of i draws, j of them minus,

(3) $v(i,j)$ is the value under optimal play of the chance to continue drawing from the urn, if i balls have already been drawn, j of them minus.

We observe that $0 \leq v(i,j) \leq n-i$, since $n-i$ is the maximum number of plus balls remaining, and since under optimal play we will not draw if the expected value would be negative.

LEMMA 2: (1) $Q(i,j) = \binom{i}{j} q(i,j) = \dfrac{i!}{j!(i-j)!} q(i,j)$;

(2) *If i balls have been drawn, j of them minus, then the probability that the $(i+1)^{st}$ ball is plus is $q(i+1,j)/q(i,j)$ and the probability of a minus is $q(i+1,j+1)/q(i,j)$*;

(3) $q(i,j) = q(i+1,j) + q(i+1,j+1)$.

PROOF: Part (1) is the consequence of there being $\binom{i}{j}$ distinct sequences of j bad and $i-j$ good balls, each sequence equally likely by Lemma 1. For part (2) let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_i)$ be a sequence of i draws, j of them minus, with each $\alpha_k = -1$ or $+1$. Then the probability of α is $q(i,j)$. Let Q^+ be the probability that the next ball drawn is plus. Then the probability of the sequence $\alpha' = (\alpha_1, \ldots, \alpha_i, +1)$ is $q(i,j)Q^+$. But α' is a sequence of $i+1$ draws, j of them minus, so its probability is $q(i+1,j)$. Thus $q(i+1,j) = q(i,j)Q^+$ and $Q^+ = q(i+1,j)/q(i,j)$. Similarly $Q^- = q(i+1,j+1)/q(i,j)$ where Q^- is the probability of a minus. Part (3) follows from part (2) since $Q^+ + Q^- = 1$.

We are now prepared to construct our strategy for the $(n;P)$ urn. The definition of the strategy will not require any consideration of Q, q, or v. Instead, the optimal strategy will be based on quantities $b(i,j)$, which are computed using auxiliary quantities $a(i,j)$ derived directly from n and P. The optimality of the strategy is due to the relationships

$$b(i,j) = q(i,j)v(i,j), \quad a(i,j) = q(i+1,j) - q(i+1,j+1).$$

THEOREM 4: *Given an $(n;P)$ urn, for $0 \leq j \leq n$ let $P^*(j) = P(j)/\binom{n}{j}$ and $a(n,j) = b(n,j) = 0$. For $0 \leq j \leq n-1$ let $a(n-1,j) = P^*(j) - P^*(j+1)$, and for $0 \leq i \leq n-2$ and each j let*

$$a(i,j) = a(i+1,j) + a(i+1,j+1).$$

For $0 \leq i \leq n-1$ and each j let

$$b(i,j) = \max\{0, a(i,j) + b(i+1,j) + b(i+1,j+1)\}.$$

Then an optimal stopping rule for maximizing the expected score at the stopping time is: if i balls have been drawn, j of them minus, draw again only if $b(i,j) > 0$. The value $V(n;P)$ of the $(n;P)$ urn under optimal play is $b(0,0)$.

PROOF: We first show that

$$a(i,j) = q(i+1,j) - q(i+1,j+1).$$

This is true for $i = n-1$ since $P(j) = Q(n,j)$ and therefore $P^*(j) = q(n,j)$. For $i \leq n-2$ the formula for $a(i,j)$ follows by induction using part (3) of Lemma 2. We next show that $b(i,j) = q(i,j)v(i,j)$. Since $v(n,j) = 0$ by definition, we have $b(n,j) = 0 = q(n,j)v(n,j)$. For $i < n$, we call on the principle used in Theorem 5, so $v(i,j)$ satisfies

$$v(i,j) = \max\left\{0, \frac{q(i+1,j+1)}{q(i,j)}(-1 + v(i+1,j+1)) + \frac{q(i+1,j)}{q(i,j)}(+1 + v(i+1,j))\right\}$$

since by part (2) of Lemma 2 the ratios of the q's are the transition probabilities. Since $q(i,j) \geq 0$ and $x \cdot \max\{0,y\} = \max\{0,xy\}$ when $x \geq 0$, we can multiply through by $q(i,j)$ and use the formula for $a(i,j)$ to obtain

$$q(i,j)v(i,j) = \max\{0, a(i,j) + q(i+1,j)v(i+1,j) + q(i+1,j+1)v(i+1,j+1)\}.$$

Then using the definition of $b(i,j)$, the identity of $b(i,j)$ and $q(ij)v(i,j)$ follows by induction.

It is clear that an optimal stopping rule is to stop whenever $v(i,j) = 0$. We next show that stopping when $b(i,j) = 0$ is equivalent. If $q(i,j) > 0$, then $v(i,j) = 0$ if and only if $b(i,j) = 0$; while if $q(i,j) = 0$, then $Q(i,j) = 0$, so the probability that j of the first i balls drawn will be minus is zero, and the (i,j) choice will never occur.

Thus the value of $b(i,j)$ is irrelevant when $q(i,j) = 0$, and a strategy based on $b(i,j)$ is optimal. To obtain $V(n;P)$, we note that

$$Q(0,0) = 1\,, \quad \binom{0}{0} = 1\,,$$

so

$$q(0,0) = 1$$

and

$$b(0,0) = q(0,0)v(0,0) = v(0,0) = V(n,P)\,.$$

Appendix 3

A property of optimal strategies for random urns

■ This theorem shows that although the strategies for the random urn can be quite varied, there are some common characteristics.

THEOREM 7: *Given an $(n;P)$ urn, suppose that i balls have been drawn, j of them minus, and that $b(i,j) > 0$. Then under optimal play the probability that all balls will be drawn is positive.*

PROOF: We show that if $b(i,j) > 0$, then either $i = n - 1$, $b(i+1,j) > 0$, or $b(i+1,j+1) > 0$. Then since $b(i+1,j) > 0$ implies $Q^{+} = q(i+1,j)/q(i,j) > 0$, and $b(i+1, j+1) > 0$ implies $Q^{-} = q(i+1,j+1)/q(i,j) > 0$, with positive probability play will continue. Note that by definition $b(k,l) \geq 0$ so $b(k,l) \geq a(k,l)$. Then if $i < n - 1$ and both $b(i+1,j) = 0$ and $b(i+1,j+1) = 0$, we must have $a(i+1,j) \leq 0$ and $a(i+1,j+1) \leq 0$. But then from the definition of $a(i,j)$,

$$a(i,j) = a(i+1,j) + a(i+1,j+1) \leq 0\,,$$

so $b(i,j) = \max\{0, x + y + z\}$ where neither x, y, nor z exceeds zero. But then $b(i,j) = 0$, contradicting the hypothesis, so either $i = n - 1$, $b(i+1,j) > 0$, or $b(i+1,j+1) > 0$.

Other properties of random urns are discussed in [2].

Appendix 4

Stopping-time statistics

■ In this section we present recursion formulae for the joint distribution of stopping time and stopping-time score. Although we do not use these in this paper, the study of the statistics of the stopping time (or "holding period") and score appears to be an interesting area for research.

THEOREM 8: *Given an $(n;P)$ urn, set $S(i,j) = 0$ for pairs with $b(i,j) > 0$. For pairs with $b(i,j) = 0$, let $S(i,j)$ be the number of sequences of draws which end at (i,j) without passing through any (i',j')* for which $i' < i$ *and $b(i',j') = 0$. Then the probability that under optimal play stopping will occur at (i,j), that is, with i balls drawn, j of them minus, is $S(i,j)q(i,j)$.*

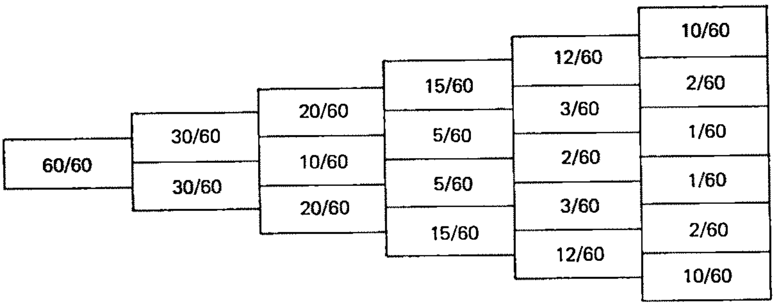
We compute $q(i,j)$ from part (3) of Lemma 2, starting with $q(n,j) = P^{*}(j)$. For our equiprobable example with $n = 5$, $q(i,j)$ is given in Figure 17. To compute $S(i,j)$ we use the next theorem.

THEOREM 9: *$S(i,j)$ may be computed by the following algorithm: set $c'(0,0) = 1$; if $c'(i,j)$ has been defined, define $c(i,j)$ and $S(i,j)$ by*

(1) if $b(i,j) > 0$, set $c(i,j) = c'(i,j)$ and $S(i,j) = 0$;
(2) if $b(i,j) = 0$, set $c(i,j) = 0$, and $S(i,j) = c'(i,j)$.

FIGURE 17

q (i,j)

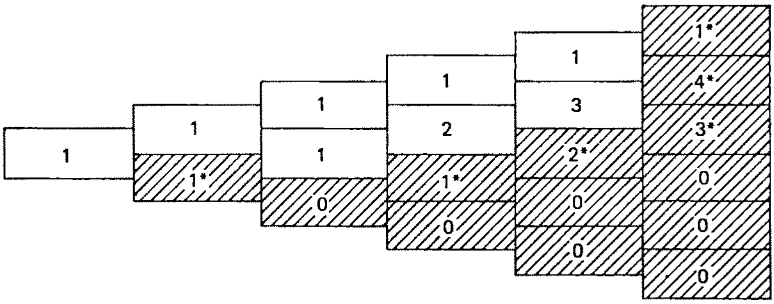


Continue by defining $c'(i+1,i+1) = c(i,i)$, $c'(i+1,0) = c(i,0)$, *and* $c'(i+1,j) = c(i,j) + c(i,j-1)$ *when* $0 < j < i+1$.

Figure 18 illustrates the computation of $S(i,j)$ using Theorem 9 for our equiprobable example. Those squares with $b(i,j) = 0$ are

FIGURE 18

S (i,j) SHADED AND c (i,j) UNSHADED

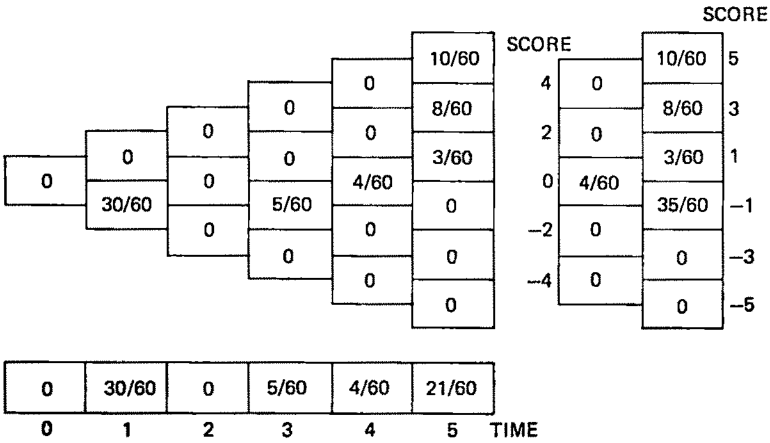


shaded. The figures in the shaded squares are $S(i,j)$; those in the unshaded squares are $c(i,j)$. The only squares on which stopping occurs are the shaded squares with $S(i,j) > 0$; for those, the numbers are starred.

In Figure 19 we tabulate $S(i,j)q(i,j)$ along with the marginal distributions of stopping time (i) and score $s = (i-j) - j = i - 2j$.

FIGURE 19

S (i,j) q (i,j) AND MARGINALS



For this example, the stopping time has mean 83/30 = 2.77. and standard deviation 1.85, while the score has mean 42/60 (which checks) and standard deviation 0.96.

Appendix 5

Two-point strategy plots

■ To generalize Shepp's Theorem 1 in a different direction, we have also investigated the nature of optimal strategies for two-point distributions, $P(m) = 0$ except for $m = m_1$, m_2. Given m_1 and m_2, $P(m_1)$ and $P(m_2)$ are chosen so that $E(P) = n/2$. Typical cases are shown in Figures 20 and 21. In Figure 20 we see that a string of losses will cause us to quit, since we are probably heading for the lower value, but if we are enough higher than the higher final value we will quit also. Under optimal play we will never get into the lower playing set, since to get there we would have to pass through a stopping area (·). But if by mistake we got into the lower block of X's, we should keep playing until hitting a (·). Figure 21 is similar, except that an initial string of losses will lead us into the lower playing set even under optimal play. Note that in both figures the "boundary function" is not single-valued as it was in the Gaussian cases.

FIGURE 20

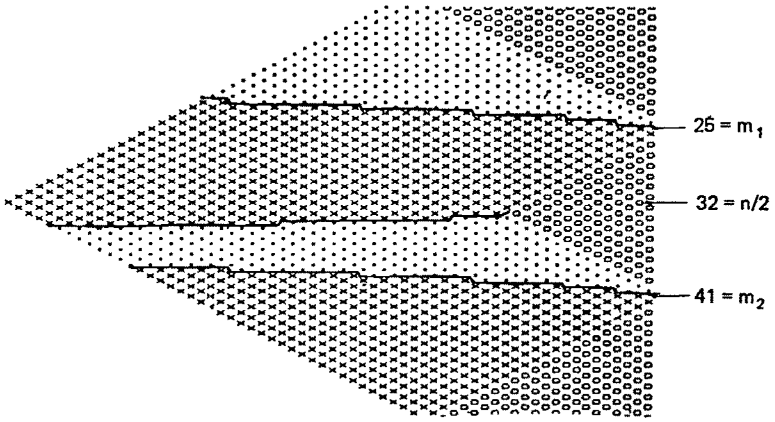
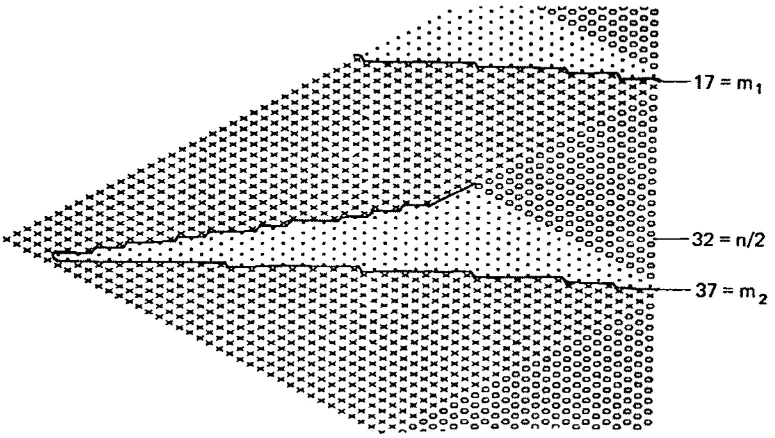


FIGURE 21



A two-point FFT distribution such as these might be appropriate for situations involving changes in laws or Government regulations, or perhaps for currency revaluations. However, a basic process other than the Wiener process, perhaps a discontinuous one, might be preferable for modeling these situations.

References

1. Bachelier, L. "Theorie de la Speculation," *Ann. Sci. École Norm. Sup.* (3), No. 1018, Gauthier-Villars, Paris, 1900. Reprinted in translation in [4].
2. Boyce, W. M. "Optimal Drawing from a Random Urn." In preparation.
3. Breiman, L. "Stopping-Rule Problems," in *Applied Combinatorial Mathematics.* New York: Wiley, 1964.
4. Cootner, P. H., ed., *The Random Character of Stock Market Prices.* Cambridge: MIT Press, 1964.
5. Feller, W. *An Introduction to Probability Theory and its Applications.* Vol. 1, 2nd ed. New York: Wiley, 1957.
6. Granger, C. W. J. and Morgenstern, O. "Spectral Analysis of New York Stock Market Prices," *Kyklos*, 16 (1963), pp. 1–27. Reprinted in [4].
7. Shepp, L. A. "A First Passage Time for the Wiener Process," *The Annals of Mathematical Statistics*, Vol. 38, No. 6 (December 1967), pp. 1912–14.
8. ———. "Explicit Solutions to Some Problems of Optimal Stopping," *The Annals of Mathematical Statistics*, Vol. 40, No. 3 (June 1969), pp. 993–1010.

Regulation and the durability of goods

Richard Schmalensee

Assistant Professor, Sloan School of Management,
Massachusetts Institute of Technology

This paper considers the production of durable goods that deteriorate at a constant percentage rate under conditions of competitive and monopoly equilibrium. A perfect market in used units of the good is assumed, as is constant returns to scale in production. It is shown that a monopolist will generally produce goods that deteriorate faster than those produced by a competitive industry. Regulation of the durability of the monopolist's output will always cause him to increase the volume of services of the good available to the market. Government intervention to limit the monopolist's price-cost margin may cause the firm to increase the deterioration rate of its output, and such regulation may also lead the monopolist to decrease the volume of services of the good available to the market. In this case, regulation will have effects exactly contrary to the presumed intent of the regulators. The impact of price regulation is shown to depend on the shape of the demand curve for the services of the durable good. Finally, it is pointed out that both price and durability must be regulated if a monopolist is to be made to reproduce competitive performance.

■ Recent research has established some conditions under which a monopolist would produce goods of less durability than a competitive industry. The principal paper is Levhari and Srinivasan [1], hereinafter LS. That work considers goods which do not deteriorate in use. After a certain period of time, however, the goods abruptly evaporate and cease to provide any services. These are termed "one-hoss shay" goods. LS found that a monopolist will generally produce one-hoss shay goods which evaporate sooner than the goods which would be produced by a competitive industry.

This result should be re-examined for other sorts of durable goods. The most common model of producer or consumer durables assumes an infinite lifetime with deterioration at a constant percentage rate. This is the so-called "radioactive decay" model. Such goods never evaporate in the one-hoss shay sense, but their ability to provide services steadily diminishes. We shall extend the LS results to this case.

Then there are the very important questions of the effects of regulation on durability and price. What if durability alone is prescribed by the government? We establish the LS result that durability regulation is generally in the public interest, since the price of a unit

---

The author obtained the S. B. degree in 1965 and the Ph D. degree in February 1970, both from M.I.T. He has served first as an Instructor and currently as Assistant Professor at M.I T.'s Sloan School of Management, and will become Assistant Professor of Economics at the University of California, San Diego. His present interests include using the computer in the teaching of economics, the economics of advertising, and microeconomics, both pure and applied.

of services of the durable good will be lowered and the volume of services available increased. We also show, however, that regulation of durability will not lower price to the competitive level. What if price alone is regulated? It is proved that restricting the allowable markup on the good may serve to increase or decrease the durability of a monopolist's output. Price regulation, by itself, will fail to induce the production of goods with the same durability as would be produced under pure competition. We show that there is no guarantee that price regulation alone will increase the volume of services available to the market. Thus, to force a monopoly to reproduce competitive performance, both price and durability must be regulated.

The essential reasoning behind this conclusion is quite simple. The demand for a durable good depends upon both initial purchase price and durability, since both influence the cost of utilizing the services of the good. If only one of these quantities is controlled, a monopolist will exploit the finite elasticity of the demand curve he faces by manipulating the other.

The method of this paper, like that of LS, is comparative statics; we shall compare long-run equilibria. The first section outlines some common features of all such equilibria. The next section extends the basic LS results to the case of deteriorating durable goods. The final section examines the consequences of regulating the behavior of a monopolist producing such goods.

■ We assume that the durable good in question is valued only for the services it provides. The ability of any unit of the good to provide services decays exponentially at a rate γ, though the good itself is assumed to last forever. (The steel in an automobile rusts rather slowly, but the utility of owning any particular automobile declines fairly rapidly over time, even in the absence of model changes.) One new unit of the good provides one unit of services, but a unit produced t years ago yields only $e^{-\gamma t}$ units of services.

We consider only situations of long-run equilibrium, with production taking place at a constant rate of y units per year. The total volume of services available to the market depends both on y and on the durability of the good, since this determines the volume of services provided by past production. We integrate over all past production to determine the total volume of services available, remembering to count only the fraction of each year's production that has not deteriorated:

$$\text{total services} = \int_0^\infty y e^{-\gamma t} dt = y/\gamma .$$

In a discrete-time model, this result is easily obtained as follows. New goods provide y units of services. Goods produced last year have deteriorated and yield only $y(1-\gamma)$ units, and goods produced t years ago provide $y(1-\gamma)^t$ units. Summing these terms from t equals zero to infinity yields the above expression for total services.

We assume a perfectly competitive market in used units of the durable good, with no transaction costs present. In this situation, the market will determine the rental rate, the price per unit of services per unit time. Any owner of the good who decides to consume its services himself is essentially renting to himself; the cost of a unit of services is the foregone rental value. The demand relation therefore expresses total services available as a function of the rental rate. We denote the rental rate as p, and we write the demand relation as

$$y/\gamma = f(p), \quad \text{with} \quad f'(p) = \frac{df}{dp} < 0. \tag{1}$$

Given the rate of deterioration, γ, and the market interest rate, r, there is a simple relation between the rental rate that will prevail and the purchase price of the good. Under certainty, the price of a new unit of the good must equal the present discounted value of the rents that could be earned by renting the good on the market forever. In equilibrium, any owner of the good must be indifferent between renting the good out forever or selling it outright, since these are essentially the same transaction. Remembering that a unit t years old provides only $e^{-\gamma t}$ units of services and hence can yield only $pe^{-\gamma t}$ in rentals, we can express this basic equality as

$$P(\gamma) = \int_0^\infty (pe^{-\gamma t})e^{-rt}dt = \frac{p}{r + \gamma}, \tag{2}$$

where $P(\gamma)$ is the purchase price of a new unit of the commodity with decay rate γ.

A simple way to obtain this same expression in the context of a discrete-time model is the following. A one-year-old unit of the good is exactly equivalent under our assumptions to $(1 - \gamma)$ units of new production. It must therefore sell for $P(\gamma)(1 - \gamma)$. A user of the services of this good must, in equilibrium, be indifferent between renting a unit of services and holding a new unit of the good for one period. Suppose that rentals are paid at the end of each period. The cost of renting is simply p, while the cost of ownership is $rP(\gamma) + \gamma P(\gamma)$. The first term is the interest earnings foregone by having $P(\gamma)$ tied up in the good, and the second component is the loss in the good's market value that the owner bears. Equating the cost of owning and the cost of renting yields equation (2) immediately.

We assume that production takes place under constant returns to scale. The average and marginal cost of producing a product with decay rate γ will be written $C(\gamma)$. Constant returns to scale are usually assumed in theoretical discussions of competition and general equilibrium. This is a natural assumption in the absence of externalities. Besides, once it is dropped, we have the problem that the competitive and monopoly situations are not directly comparable since one enjoys scale economies to begin with. If there are increasing returns, for instance, a monopoly can perform better than a competitive industry, in the sense of greater production, and still earn excess profits.

We naturally assume that it costs more per unit to produce a more durable product. This means that the first derivative of $C(\gamma)$ will be negative, since increasing γ lowers durability and makes the product less expensive to produce.

All the assumptions and derivations of this section have direct counterparts in the LS paper. The difference is that LS measure durability by the good's lifetime, while we assume infinite lifetime and are concerned with the deterioration or decay rate. They, too, assume a perfect market in used goods and their services, and constant returns to scale in production.

■ We first consider competitive equilibrium. In long-run equilibrium, marginal cost will be equal to price and profits will be zero. That is, no producer has an incentive to enter or leave the industry, and none has any incentive to change the volume of production at the "going" durability. This is the usual textbook description of competitive equilibrium; in our notation it becomes

$$C(\gamma) = P(\gamma) = \frac{p}{r+\gamma}. \tag{3}$$

The second condition for equilibrium is that no firm have an incentive to change the durability of the goods it produces. All must choose γ so as to maximize profits, even though profits will be zero in equilibrium. Since each firm takes p as given, the first-order condition for optimal γ is

$$\frac{d}{d\gamma}[P(\gamma) - C(\gamma)] = 0, \quad \text{or} \quad C'(\gamma) = \frac{-p}{(r+\gamma)^2}. \tag{4}$$

The second-order condition for a maximum reduces to

$$C''(\gamma) > \frac{2p}{(r+\gamma)^3}. \tag{4a}$$

This condition corresponds to equation (4a) in the LS paper, though our condition is more restrictive. We not only require that the marginal cost of increasing durability be rising, we require that it increase sufficienctly rapidly.

Dividing (4) by (3), we obtain

$$-\frac{C'(\gamma)}{C(\gamma)} = \frac{1}{r+\gamma}, \tag{5}$$

the solution to which is the value of the decay rate γ determined in a competitive market. In order for γ to be positive, it is necessary that the left-hand side of equation (5) be less than $1/\gamma$. That is, the elasticity of marginal cost with respect to γ must be greater than minus one at equilibrium. Notice that the decay rate does not depend on demand conditions; it is completely determined by the rate of interest and the unit cost function. This was noted by LS.

In the usual model of a competitive industry with free entry and no externalities, equilibrium market price is equal to the minimum of firms' average cost curves, independent of demand conditions. Here we observe a generalization of that result. Given γ from equation (5) p can be determined from either (3) or (4) without reference to demand conditions. Given γ and p, equation (2) gives the purchase price of the good, again without reference to the demand function. Thus durability and purchase price, as well as the rental rate, are independent of demand. Here, as in the usual competitive model, the location of the demand curve affects only the volume of production.

Let us now consider the effect of changing the interest rate. The derivative of γ with respect to r in equation (5) can easily be shown to be

$$\frac{\partial\gamma}{\partial r} = \frac{p}{C''(\gamma)(r+\gamma)^3 - 2p},$$

which is positive if condition (4a) is satisfied. We thus have the LS result that an increase in the rate of discount leads a competitive industry to produce goods of lower durability (greater decay rate).

■ We now turn to an analysis of monopoly behavior. For simplicity, we invert the demand relation to obtain

$$p = g(y/\gamma). \tag{1'}$$

The profits of the monopolist in each period are given by

$$R = y[P(\gamma) - C(\gamma)] = \frac{yg(y/\gamma)}{r+\gamma} - yC(\gamma). \tag{6}$$

Differentiating (6) with respect to y, we obtain the first necessary condition for maximum profit:

$$C(\gamma) = \frac{p}{r+\gamma}\left[1 + \frac{1}{E}\right], \quad \text{where} \quad E = \frac{g(y/\gamma)}{g'(y/\gamma)(y/\gamma)}. \tag{7}$$

The quantity E is thus the elasticity of demand for services of the durable good with respect to the price of those services, the rental rate. Equation (7) is the usual condition for optimal output of a monopoly. Recall that E must be less than minus one at monopoly equilibrium, since when demand is inelastic it is always profitable to reduce output.

Differentiation of equation (6) with respect to γ yields the second necessary condition[1]:

$$C'(\gamma) = -\frac{p}{r+\gamma}\left[\frac{1}{r+\gamma} + \frac{1}{\gamma E}\right]. \tag{8}$$

Dividing (8) by (7), we obtain

$$-\frac{C'(\gamma)}{C(\gamma)} = \frac{E}{E+1}\left[\frac{1}{r+\gamma} + \frac{1}{\gamma E}\right]. \tag{9}$$

Our equations (9) and (5) correspond to the LS equations (10) and (5).

In the monopoly case, all equilibrium quantities are influenced by demand conditions. The four equations (7), (8), (1), and (2) must be solved simultaneously to determine γ, p, $P(\gamma)$, and y. This, too, corresponds to the usual textbook result.

Following the LS line of proof, we can show that the right-hand side of equation (9) is always less than the right-hand side of (5):

$$\frac{E}{E+1}\left[\frac{1}{r+\gamma} + \frac{1}{\gamma E}\right] - \frac{1}{r+\gamma} = \frac{1}{r+\gamma}\left[\frac{E}{E+1} - 1\right] + \frac{1}{\gamma(E+1)}$$

$$= \frac{-1}{E+1}\left[\frac{1}{r+\gamma} - \frac{1}{\gamma}\right] < 0,$$

---

[1] Differentiating equation (6) directly and dividing by y, we obtain

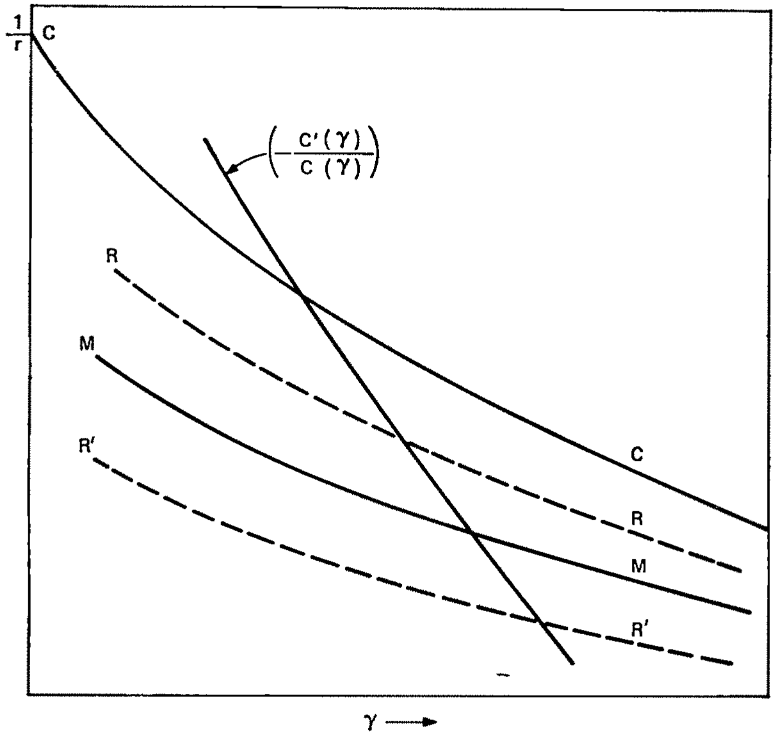
$$\frac{(r+\gamma)g'(y/\gamma)(-y/\gamma^2) - g(y/\gamma)}{(r+\gamma)^2} - C'(\gamma) = 0.$$

This can be re-arranged to yield

$$C'(\gamma) = -\frac{g(y/\gamma)}{(r+\gamma)}\left[\frac{1}{r+\gamma} + \frac{g'(y/\gamma)(y/\gamma)}{g(y/\gamma)\gamma}\right].$$

Employing equation (1′) and the definition of E, we obtain equation (8).

FIGURE 1



since r and γ are positive, and E must be less than minus one. Since the right-hand side of equation (5) falls with γ, the situation is as depicted in Figure 1. There the curve *CC* is the right-hand side of (5), corresponding to competitive equilibrium, and *MM* is the right-hand side of (9), corresponding to monopoly equilibrium.

We now show that $[-C'(\gamma)/C(\gamma)]$ intersects the competitive curve *CC* from above:

$$\frac{d}{d\gamma}\left[-\frac{C'}{C}-\frac{1}{r+\gamma}\right]=-\frac{C''}{C}+\left[\frac{C'}{C}\right]^2+\left[\frac{1}{r+\gamma}\right]^2$$

$$=-\frac{C''(r+\gamma)}{p}+\left[\frac{2}{r+\gamma}\right]^2<-\left[\frac{2}{r+\gamma}\right]^2+\left[\frac{2}{r+\gamma}\right]^2=0.$$

where conditions (3), (4), and (4a) have been employed. The second-order conditions for monopoly equilibrium do not seem to yield such a clean result, since the decisions on γ and y are interdependent. Nevertheless, if $[-C'/C]$ is a monotonic decreasing function, it will always intersect *MM* from above also, as shown in Figure 1. Since γ is larger at the intersection of $[-C'/C]$ with *MM* than at the intersection with *CC*, we have the *LS* result that a monopoly will produce goods with lower durability.

We can illustrate this conclusion with a simple example. Suppose that the demand function exhibits constant price elasticity:

$$y/\gamma = f(p) = Ap^E,$$

where E is less than minus one. Assume also that the unit cost function has constant elasticity with respect to γ:

$$C(\gamma) = B\gamma^{-b},$$

where b is a positive constant. Substituting into equation (5), it is easy to show that the competitive decay rate is given by

$$\gamma_c = \frac{br}{1-b}.$$

For this to be positive, we must have b less than one, as we mentioned below equation (5). As long as b is greater than $(\gamma - r)/(\gamma + r)$, condition (4a) is satisfied. Thus there is always a range of admissible values for b. The above expression illustrates the positive relation between r and γ, as well as the independence of γ from demand conditions.

Substitution into equation (9) yields the monopoly decay rate:

$$\gamma_m = \frac{br}{1-b} \frac{(E+1)-(1/b)}{(E+1)}.$$

From these two expressions, it should be clear that γ_m is always greater than γ_c, and the monopolist will always produce less durable goods.

■ We now turn to an examination of the effects of regulation of a monopolist producing a durable good, by setting standards for durability and then by price control. In the usual textbook model the measure of a firm's or industry's performance is the volume of output, which is inversely related to the price charged. Here the chief measure is the volume of services available to the market. Total services available will be inversely related to the rental rate.

Let us first analyze the case where the decay rate, γ, is fixed by the government. Equation (6) still describes the monopolist's profit, but now the firm's only decision variable is y, the volume of production. The necessary condition for maximum profits is still equation (7), which we re-write as

$$\frac{dR}{dy} = F(\gamma,y) = \frac{g(y/\gamma) - (y/\gamma)g'(y/\gamma)}{(r+\gamma)} - C(\gamma) = 0. \qquad (10)$$

The second-order condition for a maximum is

$$\frac{\partial F}{\partial y} = -\frac{yg''(y/\gamma)}{\gamma^2(r+\gamma)} < 0. \qquad (10a)$$

This is equivalent to requiring g'' and f'' to be positive. In words, the demand function must be convex to the origin.

We first consider the effect of durability regulation on y. If decreasing the prescribed γ always caused the firm to raise y, then it would be certain that the volume of services, y/γ, would be increased by stricter regulation. By the implicit function theorem, we have

$$\frac{dy}{d\gamma} = -\frac{\partial F}{\partial \gamma}\bigg/\frac{\partial F}{\partial y},$$

which, by the second-order condition given above, has the sign of $(\partial F/\partial\gamma)$. Differentiating equation (10) with respect to γ and employing (10) and (10a) to substitute for $g'(y/\gamma)$ and $g''(y/\gamma)$, we obtain

$$\frac{\partial F}{\partial \gamma}\frac{1}{C(\gamma)} = \left[-\frac{C'(\gamma)}{C(\gamma)} - \frac{1}{r+\gamma}\right] - \left[\frac{y}{C(\gamma)\gamma}\frac{\partial F}{\partial y}\right]. \tag{11}$$

We showed above that the first expression in brackets falls with γ near the competitive equilibrium. It is zero at that point, from equation (5), so it must be negative for γ larger than the competitive value. By the second-order condition above, however, the second bracketed term on the right of equation (11) is negative. Whether y rises or falls as γ is decreased thus depends on the relative magnitude of the two terms; it can go either way.

Consider a regulation requiring the monopolist to reduce the deterioration rate of the goods he produces by one percent. If he obeys this order and does not change production, the volume of services available to the market will clearly rise. But it may be profitable to lower y in the face of stricter durability standards. The volume of services available will increase even if he lowers y, as long as he lowers y by less than one percent. We are thus led to consider the elasticity of y with respect to γ, which we shall call ω. If ω is less than one, forcing the monopolist to increase durability will cause him to increase the volume of services available to the market. Multiplying equation (11) by $[-\gamma C(\gamma)]$ and dividing by $[y\partial F/\partial y]$, we obtain directly

$$\omega = \left[-\frac{C'(\gamma)}{C(\gamma)} - \frac{1}{r+\gamma}\right]\left[\frac{-\gamma C(\gamma)}{y(\partial F/\partial y)}\right] + 1. \tag{11a}$$

We established above that the first term in brackets on the right of (11a) is negative for γ above the decay rate that would be produced at competitive equilibrium. The right-most bracketed term is easily seen to be positive by virtue of equation (10a). The product of these two expressions is thus always less than zero. Since it states that ω is equal to one plus a negative number, equation (11a) clearly establishes the fact that ω is less than one. We thus have the result that as long as the monopolist is not required to produce goods more durable than those that would be produced by a competitive industry, the greater the required durability the greater the volume of services that will be available to the market. And, via the demand relation, the rental rate will be lowered by tightening durability standards. This same conclusion was reached by LS for one-hoss shay goods, though their line of proof differs from ours.

Equations (3) and (10) may be re-written as follows:

$$P(\gamma) = C(\gamma), \tag{3'}$$

$$P(\gamma) = C(\gamma)\left[\frac{E}{E+1}\right]. \tag{10'}$$

If the monopolist is required to produce goods with the competitive decay rate, the value of γ will be the same in both these equations. But it is clear that because the elasticity of the monopolist's demand curve is less than infinite, he will charge a price above the competitive level. From equation (2), this is equivalent to saying that the price of services, the rental rate, will be above the competitive level. And from the demand relation, equation (1), the total volume of services on the market will be smaller. Durability regulation alone cannot force the monopolist to reproduce competitive performance. This same result holds for one-hoss shay goods, as may be seen by comparing equations (3) and (17) in LS.

Now, suppose that the authorities regulate the price policy of the monopolist. In particular, assume that the ratio of sale price to unit cost, α, is fixed between one and the value of $E/(E+1)$ that prevails at the unregulated monopoly equilibrium. The monopolist will now choose γ so as to maximize his profit subject to the regulatory constraint

$$P(\gamma) = \alpha C(\gamma), \quad \text{or} \quad p = \alpha(r+\gamma)C(\gamma), \tag{12}$$

recalling equation (2).

Employing equations (1) and (12), we can re-write equation (6) and express the monopolist's profits as

$$R = (\alpha - 1)\gamma f[\alpha(r+\gamma)C(\gamma)]C(\gamma). \tag{6'}$$

The first-order condition for a maximum of profits with respect to γ can be shown, after a bit of algebra, to be

$$-\frac{C'(\gamma)}{C(\gamma)} = \frac{E^*}{E^*+1}\left[\frac{1}{r+\gamma} + \frac{1}{\gamma E^*}\right]. \tag{13}$$

This is formally identical to equation (9), but the interpretation is different. In this environment, the regulatory authority determines the price elasticity, E^*, that corresponds to every γ. That is, given γ, equation (12) yields p as a function of α, and price elasticity is clearly a function only of the price, p. Condensed into a formula, this relation becomes

$$E^* = \frac{f'[\alpha(r+\gamma)C(\gamma)][\alpha(r+\gamma)C(\gamma)]}{f[\alpha(r+\gamma)C(\gamma)]}. \tag{14}$$

It is easy to show that the right-hand side of (13) falls as E^* is increased for any E^* and γ. The curve RR shows the right-hand side of (13) when demand is made more elastic at all values of γ. That is, E^* is less than E for all γ. Regulation has brought about greater durability. When regulation decreases price elasticity, a curve like $R'R'$ results, and the decay rate is higher with price regulation than without. There is no change in durability if the price elasticity of demand is unaltered by regulation. This can happen either if the regulatory constraint is not binding or if $f(p) = Ap^E$. In this latter case, $E = E^*$ and the curve MM continues to apply.

We are thus interested in the conditions under which a fall in α will lower E^* as a function of γ, as this is when stricter price regulation will always increase durability. From (14), however, E^* will fall when α falls if and only if it will fall when p falls. This is because stricter price regulation lowers the rental rate corresponding to every value of γ by lowering purchase price. If, on the other hand, $f(p)$ becomes less elastic as p falls, E^* will be greater than E, and regulating a monopolist's markup will lead him to lower the durability of his product. The condition that E^* must vary directly with p at any particular point can be written formally as

$$\eta > 1 - E, \quad \text{where} \quad \eta = -\frac{f''(p)p}{f'(p)}. \tag{15}$$

The quantity η is thus minus the elasticity of $f'(p)$ with respect to p. It will have the sign of $f''(p)$.

It is of some interest to note that a number of commonly encountered demand curves fail to satisfy equation (15) anywhere. When E is constant, it is easy to show that $\eta = 1 - E$ and durability is unchanged, as we asserted above. In the case of each of the following demand functions, it can be shown that η is less than $1 - E$ and therefore stricter price regulation will cause durability to decrease:

$$(y/\gamma) = a - bp,$$
$$(y/\gamma) = a - b\ln(p),$$
$$\ln(y/\gamma) = a - bp,$$

where a and b are positive constants. All of these functions are convex to the origin, with $f''(p)$ greater than or equal to zero. But $f''(p)$ is not large enough to cause (15) to be satisfied. Condition (15) will, of course, never be satisfied by a demand function for which $f''(p)$ is negative. I have been unable to find any examples in the literature of demand functions that satisfy condition (15). We thus have the result that stricter price regulation is likely to cause a monopolist to produce goods with a larger deterioration rate.

A natural question at this point is whether or not price regulation will increase the volume of services available, as durability regulation will. From equation (1), the volume of services will increase if the rental rate falls. Differentiating equation (12) and dividing, we obtain

$$\frac{dp}{d\alpha}\frac{1}{(r+\gamma)C(\gamma)} = 1 + \alpha\left[\frac{C'(\gamma)}{C(\gamma)} + \frac{1}{r+\gamma}\right]\frac{d\gamma}{d\alpha}. \tag{16}$$

For stricter price regulation to increase the volume of services available, the right-hand side of (16) must be positive, since lowering α must also lower p. For γ above the competitive equilibrium value, the term in brackets is positive, as we showed below equation (11). Thus if lowering the allowable markup also lowers γ, the volume of services will be increased. On the other hand, if lowering α causes less durable goods to be produced, as we saw above is likely, then equation (16) shows that we have no guarantee that the volume of services will increase.

Finally, it is clear that E^* will be finite for all γ and α. That is, the right-hand side of (13) will always be less than the right-hand side of (5). This means that the curve *RR* will always lie below *CC* in Figure 1. Even in the favorable case where demand is made more elastic by regulation, it is impossible to induce a monopolist to produce goods as durable as those that would be produced by a competitive industry by regulating only his pricing policy.

In this section we have shown that neither fixing the price-cost margin nor setting standards for durability alone will force a monopolist to reproduce competitive performance. If durability standards are set, the monopolist will increase the volume of services available to the market, though he may well lower his production as measured in units of the good. If the price-cost margin alone is regulated, it is likely that the durability of the monopolist's output will decline, and it is possible that the volume of services available to the market will also be restricted. In both cases the monopolist uses the unregulated policy variable to act against the intent of the regulators. Unless both price and durability are regulated, a rational monopolist will exploit the fact that the elasticity of demand for the services of the good he produces is finite. Regulation of price policy alone may actually cause the price of services of the good to increase, imposing a clear burden on purchasers of the good.

Reference

1. Levhari, D. and Srinivasan, T. N. "Durability of Consumption Goods: Competition Versus Monopoly," *American Economic Review*, Vol. LIX, No. 1 (March 1969).

Alternative rate of return concepts and their implications for utility regulation

Ezra Solomon

Dean Witter Professor of Finance
Graduate School of Business
Stanford University

The rate of return on invested capital is a widely used concept in both regulated and unregulated sectors of the economy. It provides a measure of actual performance as well as required or expected performance (the latter is often termed the "cost of capital"). In the utility field, regulatory agencies often focus on the rate of return as a major instrument for assessing and controlling the performance of firms under their jurisdictions. Unfortunately, two altogether distinct units are employed for measuring rate of return: (1) book rate units and (2) discounted cash flow units. Rarely will the two produce the same result, and the use of one measure as a surrogate for the other may prove highly misleading. This paper indicates the relationship between the two measures and shows the impact of some variations in depreciation and expensing procedures, growth rate, etc. The object is to point out the potential hazards associated with the use of measures of different things in a context that requires the use of measures of the same thing.

■ The rate of return on invested capital is a central concept in financial analysis. It is widely used as a basis for decisions, both in the unregulated sector and in utility regulation. One essential process for both purposes is an intercomparison between two facets of the rate of return measure:

(1) One facet is a summary measure of actual or prospective performance i.e., a measure of the annual rate at which each unit of capital input generates net financial benefits.

(2) The second facet is a summary measure of a "required" annual rate i.e., the financial standard or target rate against which the performance of prospective or already committed capital inputs can be assayed. In this second form it is also called "the cost of capital" rate.

When factor prices (other than capital) and product prices are taken as given, (as in the competitive pricing model), an intercomparison of these two facets of rate of return determines changes in the level of capital inputs. If, on the other hand, the scale of existing and future investment and factor prices (other than capital) are taken as given (as in the administered pricing model), then an intercomparison of these two facets determines product pricing decisions.

Professor Solomon holds the A.B. degree, with honors (University of Rangoon, 1940), and the Ph.D. degree (University of Chicago, 1950). He was on the faculty of the University of Chicago Graduate School of Business from 1948 to 1960, and was appointed Professor of Finance in 1957 He has been with the Graduate School of Business at Stanford University since 1961, first as Professor of Finance and Director of the International Center for the Advancement of Management Education (1961–'64), and as Dean Witter Professor of Finance since 1965.

Thus, large companies may set prices in order to achieve a target rate of return,[1] and the thrust of utility price regulation is to provide utilities an opportunity to earn some "fair" rate of return on investment.

If the "rate of return on investment" were itself a single, unambiguous concept, the only difficulties we would encounter in using it for either purpose would be difficulties involving correct estimation of the measure. But the concept is not unambiguous: quite apart from trivial variants such as pre-tax vs. post-tax measures or a total capital vs. an equity capital basis for the concept,[2] there is a non-trivial problem which arises from the fact that "rate of return" is measured in terms of two altogether distinct units: book rate units, and discounted cash flow units.

☐ **Book rate units.** These are more properly called book-ratio units. The "rate" being measured is defined as the ratio of income during a given period of time (as defined by the usual accounting measure of this term) to the net book-value of invested capital outstanding during the period (as defined by the balance sheet corresponding to the income statement from which the numerator is derived).[3] This version of the rate of return will be symbolized as *b*, and it represents the most commonly used basis for reporting and analyzing "rate of return on invested capital."[4]

☐ **DCF units.** Unlike the book-ratio, this measures the return on capital in terms of the annual rate at which the "future" (actual or prospective) net funds flows (or cash flows as these are commonly called) from an investment have a discounted value equal to the value of the investment outlays required to bring about these funds flows. Hence the name DCF units, which refers to the "discounting cash flows" process required to calculate this version of rate of return on investment.

This basis for measuring rate of return is the most commonly used one for theoretical purposes. It will be symbolized by the letter *r*.[5]

Conceptual differences between the DCF rate and the book-ratio

■ There are three major conceptual differences between the two measures, *b* and *r*.

(1) The book-ratio, *b*, defines its flow variable (income) as "cash flow" (meaning funds flow) minus depreciation and minus the expensed portion of current period investment. In contrast, the DCF rate, *r*, defines the flow variable as "cash flow" before these two adjustments.

(2) The book-ratio, *b*, defines its stock variable (investment) as the net book value of capital as this would appear on the balance

---

[1] See [9].

[2] These are trivial in the sense that they are obvious and therefore lead to no confusion. To simplify the exposition, I shall ignore these two potential differences in definition by assuming, throughout most of this paper, that income taxes and long-term borrowing do not exist.

[3] Another potential, but trivial, ambiguity arises from the fact that we can measure the accounting net book value figure on a beginning-of-period basis, an ending-of-period basis, or somewhere in between.

[4] For reporting, see *Fortune's* 500 Largest Industrials Directory, *Fortune's* 50 Largest Firms in Merchandising, Banking and Transportation; FTC-SEC Quarterly Financial Report on Rates of Return in Manufacturing, First National City Bank of New York's Annual Return on Capital series, FPC and other regulatory agencies' Annual Reports on "Return on Investment."

For analysis, see [3], [8], [11], and [18]

[5] For prospective investments it is equal to the expected marginal productivity of capital (or the marginal efficiency of capital in Keynesian analysis), or the "initial rate of return." For investments in long-term bonds it is called the "effective yield to maturity."

sheet consonant with the definition of the income variable, i.e., the balance sheet number, which is linked by the inexorable rules of double-entry accounting to the income definition. In content, the DCF rate, r, defines the stock variable (investment) as the total initial outlay of funds required for generating the cash flows counted on the flow side of the equation.

(3) Finally, the book ratio, b, defines the rate of return in a given period of time, or over a period of time, as the arithmetic ratio of its flow variable to its stock variable. In contrast, the DCF rate, r, defines the rate of return as that rate of compound discount (or interest) at which the time adjusted (present) value of the flow variable (cash or funds flows) is equal to the time adjusted (present) value of the stock variable (investment outlays).

Given these basic differences in definition between b and r, it is highly unlikely that their numerical values will be equal. Yet both carry the same label "percent per annum rate of return on investment," and the two are frequently used as if they were freely congruent and interchangeable measures of the same thing. Some examples are:

1. For a single company or industry the rate, b, is often treated as if it were an unbiased measure of r;

2. When several companies or industries are analyzed it is generally assumed that differences in b reflect corresponding differences in r;

3. Estimates of fair or reasonable rates are often calculated in DCF units, i.e., in terms of the r measure, and applied to net book value estimates without regard to the essential differences between DCF units and book rate units. Alternatively, a company may set its required rate of return for *ex-ante* capital budgeting purposes in terms of DCF cost of capital units and then measure *ex-post* performance in terms of book rate units.

These, and other forms of confusion between the two conceptually and numerically different yardsticks, can and do lead to considerable confusion in many forms of investment analysis, both in the unregulated sector and, more particularly, in utility regulation.

The rest of this paper is an attempt to explore the nature and magnitude of the differences between the r version and the b version of rate of return measurement and, on the basis of this analysis, to examine potential uses and misuses of the two concepts for interpretive and regulatory purposes.

Measuring performance: the DCF rate of return

■ For a single investment project, the DCF rate, r, is defined as the rate at which net cash flows from the project, over its productive life, have a present value equal to the original investment outlay required by the project. Thus if:

C_0 = investment outlay at time $t = 0$,
$F_t(t = 1 \cdots n)$ = net cash flows per period,
n = project life (zero salvage assumed),

we have

$$C_0 = \sum_{t=1}^{n} F_t(1+r)^{-t}. \qquad (1)$$

The DCF rate, r, is uniquely determined by the configuration of net cash inflows per unit of outlay, where configuration, in this context, refers to the volume, time-pattern, and duration of the net cash inflows F_t.[6]

In the general case, where the configuration of F_t is freely variable, the rate r has to be ascertained by trial and error calculations from compound discount tables. However for a pattern of level cash inflows, equation (1) can be restated more explicitly. Thus if the net cash inflows is a level stream of F a year for n years, we have

$$F = C_0\left[\frac{r(1+r)^n}{(1+r)^n - 1}\right]. \tag{2}$$

Regular tables exist for the bracketed items for various values of r and n, and hence the value of r for any ratio F/C_0 can be found readily by inspection. For example, assume

$$C_0 = \$1{,}000,$$
$$F = 229.61,$$
$$n = 6.$$

Then using annual end-of-period discount tables it can be shown that $r = 0.10$ or 10 percent per annum.

Although the DCF rate for a single project is a well known and widely used measure in capital analysis, the corresponding DCF measure for an ongoing company is not generally available. The reason for this is that the pattern and duration of net cash flows for a single project is known or estimated, either retrospectively or prospectively, whereas this is not true for a company, which is an ongoing collection of many projects. However, logic tells us that if a company holds many projects, each of which individually yields a DCF rate r, then the company as a whole will also be generating a DCF rate r on its total portfolio. This is true regardless of the pace at which the projects have been acquired over time and regardless of the practices used in accounting for capital and income.

To compare the results obtained by measuring return on investment in terms of the book ratio b against the underlying DCF rate r, we must therefore use the simulated model in which r itself is known.

Company profitability: *b* vs *r* in the steady state

■ In order to analyze the level of b relative to any given level of r, we take a hypothetical company or companies which acquire or can acquire only a single type of investment whose cash inflow characteristics are known. For illustrative purposes we can take the investment mentioned above: \$1,000 of outlay generates level cash inflows of \$229.61 a year for 6 years, after which the asset is scrapped at zero salvage value. The DCF rate of return on this investment is 10 percent, and hence any company which holds a portfolio consisting exclusively of such projects must be earning a DCF rate of 10 percent per annum.

---

[6] More specifically, the configuration of cash inflows per unit of outlay uniquely determines the rate r only if the outlay value is set independently of the rate, r This is clearly true in the point-input case being discussed. It also holds for investment outlays which themselves have a time duration if the discounted present value of such outlays is calculated through the use of some cost of capital rate, k which is known or assumed independently of the rate, r, expected from an individual project. However, the rate, r, is frequently (though incorrectly) calculated as that rate, r^*, that equates the present value of inflows with the present value of outlays when both inflows and outlays are discounted at a common rate, r^* For this method of computation multiple solutions for r^* can exist. For a more complete discussion of multiple solutions, see [15], ch. 10.

What book rate *b* will such a company show? As we shall see, the answer to this depends on several factors. Two factors which have a powerful effect on the size of *b* relative to *r* deserve detailed consideration. These are:

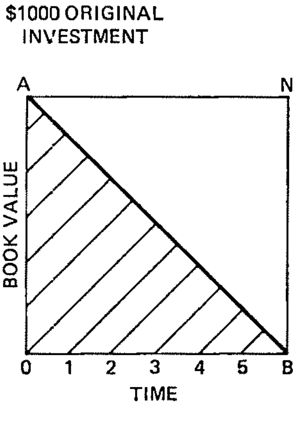
1. Accounting practices used in defining book income and net book capital, and
2. The pace at which the company acquires new investments over time.

In order to understand the effect of each of these factors, we will deal with them one at a time. Assume, to begin with, that the company acquires new investments (or projects) by investing an equal amount of money each year. This will be referred to as the "steady-state" condition (it could equally well be called the zero-growth case). Our hypothetical company which invests an equal amount each year in the basic type of project outlined will reach a "steady state" after six years. Beyond this point it will always hold six "investments." When it acquires its seventh investment, the first one it acquired is scrapped, when it acquires the eighth the second is scrapped, and so on.[7]

The book rate of return b. The book rate of return for a company in the steady state, which holds only investments identical to the basic project outlined, will depend on the accounting procedures used—in particular, on the fraction of each original investment outlay which is expensed for book purposes, and the specific depreciation formula which is used in deriving income, period by period, over each project's economic lifetime.

The book rate can be calculated by (1) defining the general value of *b* in algebraic terms and (2) solving such an equation for any given set of values of the accounting variables. However for any situation other than full capitalization of investment outlays and straight-line depreciation and zero growth, the algebraic expressions for the book rate of return can become exceedingly messy and sometimes complicated. Happily a graphical approach, even though partly intuitive, makes it possible to bypass explicit algebraic solutions and still make the relevant points.

FIGURE 1
STRAIGHT-LINE
DEPRECIATION



To begin with the simplest situation, assume that the company capitalizes all of its investment outlays and uses a straight-line depreciation formula. The income statement of such a company after it reaches a steady state will show:

Net Cash Flow = 6 × \$229.61 = \$1,377.66
Depreciation = 6 × 1/6 × \$1,000 = \$1,000.00
Net Book Income = = \$ 377.66

Its balance sheet, in the steady state, would show a net book value equal to one-half of the original outlay for the six "investments" it holds in its portfolio. This is shown in Figure 1. The original outlay is \$1,000. This is written-off continuously as time passes, i.e., the net book value of each asset is diminished along the line *AB*. After $t = 6$ the company holds six vintages of investment ranging from one 0.5

---

[7] Inflation is assumed to be zero in this portion of the analysis.

year old to one 5.5 years old, and together they have an undepreciated net book value equal to the area under the line *AB*. We can see by inspection and our knowledge of Euclid that this is $1,000 × 6/2 or $3,000.

Thus the book rate of return for this company is $377.66 ÷ $3,000 or 12.6 percent.

We now also have a tool which permits us to vary our accounting procedures and to discover the consequent effect on *b* of doing so *without* engaging in tedious algebraic expressions. (Meanwhile we still retain our basic assumption that all companies are in a "steady state," i.e., that our company acquires its investments through *equal* annual outlays.)

Variation in expensing procedures

■ We can begin our analysis by assuming again that all companies acquire only the one basic investment we have outlined ($1000 which generates $229.61 a year for six years) and that all use straight-line depreciation. However, now they are free to alter their expensing procedures, i.e., they may charge off any fraction of the $1000 as current expenses. (Recall that we are ignoring the tax-impact of expensing policies on net cash flows.)

As far as before-tax cash flows are concerned, there is no change if the company's accountant charges part of each year's $1000 outlay to current expenses rather than to book capital. Therefore accounting procedures do not affect the DCF yield *r*. It remains at 10 percent on each investment held, and hence at 10 percent for all of them collectively.

Do these accounting variations in the fraction of each year's $1000 outlay which is expensed affect the flow of company income? In the steady state (equal investment outlays each year) the answer is "No." This can be seen clearly in Table 1. Regardless of the amount expensed each year, the total annual charge of expensing plus depreciation adds up to $1000.

What Table 1 shows is that (in the steady state) the flow of reported income (after depreciation of non-expensed capital and the expensed portion of current investment) does not vary at all with the expensing policy adopted.

Although expensing policy has no effect on book income (ignoring taxes) it does have a profound effect on the net book value of capital. Returning to Figure 1, we see that with a zero expensing policy the straight-line depreciation function (Line *AB*) produces a

TABLE 1
STRAIGHT-LINE DEPRECIATION

EXPENSING POLICY	ANNUAL CASH FLOW	ANNUAL AMOUNT EXPENSED	ANNUAL DEPRECIATION	ANNUAL BOOK INCOME
0	$1377 66	$ 0	$1000	$377 66
20%	1377.66	200	800	377 66
50%	1377.66	500	500	377.66
100%	1377.66	1000	0	377.66

company net book value of $3000, i.e., the net book value figure is equal to the area under the depreciation curve (or line).

With a 100-percent expensing policy, the depreciation function will follow the right-angled line *AOB*. The area under this curve is zero. Such a company would have no net book capital. But, as we have seen in Table 1, its net income is $377.66. Its book rate of return *b* will therefore be infinity. The corresponding net book value and book rate of return for different expensing policies (straight-line depreciation) are shown in Table 2.

TABLE 2

EXPENSING POLICY	ANNUAL BOOK INCOME	COMPANY NET BOOK VALUE	b (PERCENT)
0	$377.66	$3000	12.6
20%	377.66	2400	15.4
50%	377.66	1500	25.2
90%	377.66	300	125.9
100%	377.66	0	∞

The effects of expensing policies on book rates of return are clearly powerful. What this means is that two companies which are in fact generating the same DCF rate r might in theory show book rates $b_a, b_b, b_c, \cdots b_z$ which range all the way from less than r percent at one extreme to ∞ percent at the other. The empirical question is: Do companies or divisions of companies, in fact, use different expensing ratios as far as investment outlays are concerned. The answer is that they do. Many companies capitalize all or almost all of their investment outlays. Others expense a high fraction. Companies with high research and development expenditures use a high expensing policy. Companies with high long-range advertising expenditures (i.e., expenditures which contribute little to current period cash flow but which contribute to future cash flow) do in effect use a high expensing policy. Producing departments of petroleum companies or primarily producing companies in the oil and gas industry also expense a high fraction of outlays in the form of exploratory, developmental, and intangible drilling costs.

Because our model, thus far, is confined to steady-state situations (equal or approximately equal investment outlays each year) it is premature to extrapolate our findings to the real world, but the results are suggestive. For example, the producing segment of the integrated oil industry, or oil and gas companies which are primarily producers rather than refiners and transporters, tend to have significantly higher book rates of return than the integrated companies. For example, Amerada earned an average book rate on equity capital of 21.6 percent during the period 1964–1968 as opposed to a corresponding rate of about 12.0 percent for the integrated petroleum industry as a whole. Likewise, pharmaceutical companies and cosmetic companies (which also follow a higher-than-typical expensing policy with respect to investment outlays) consistently show significantly higher rates than the rest of the manufacturing sector.

For example, according to the Fortune Directory, the pharmaceutical industry has shown the highest return on invested capital year after year. For 1968 it was again in first place with a median book rate of return of 17.9 percent (compared to the *all*-industry median rate of 11.7 percent). The soaps and cosmetics industry was in second place with a median book rate of return of 16.9 percent.

The conventional explanation for these higher-than-normal rates of return for companies or industries is that they are either:

(1) riskier (this is the standard explanation for the producing sector of the oil and gas industry),

(2) more efficient, or

(3) have monopoly powers (this is frequently applied to the pharmaceutical sector).

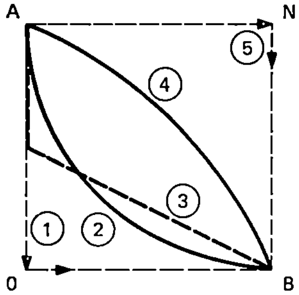
While these three conventional explanations may be correct in varying degrees, they all assume that the observable book rates accurately reflect commensurate DCF rates. The fact that many high-book-rate companies or industries also follow high "expensing" policies suggests strongly that a fourth potential explanation is too important to ignore: namely that the observable book rates significantly overstate the underlying DCF rates actually being earned.

Variations in depreciation methods

■ In addition to variations in accounting expensing policies, companies and industries use varying methods of depreciating the capitalized portion of their investment outlays. This is also a potential source of disturbance as far as the book rate of return is concerned. Since "depreciation" is a variant of "expensing" (which is instantaneous "depreciation") or vice-versa (depreciation is a form of "expensing" over time), we can deduce its effects without going through the mechanics of it in detail.

Depreciation, in itself, has no effect at all on before-tax net cash flows. Nor, in the steady state we are examining (equal annual investment outlays), does it affect net income flows. Any depreciation policy, like any expensing policy, or any combination of the two, must lead to a total annual charge equal to the annual outlay. In our arithmetic example this is $1000. In short, for any combination of depreciation policy or expensing policy, reported income in the steady state will be equal to $1377.66 less $1000, or $377.66.

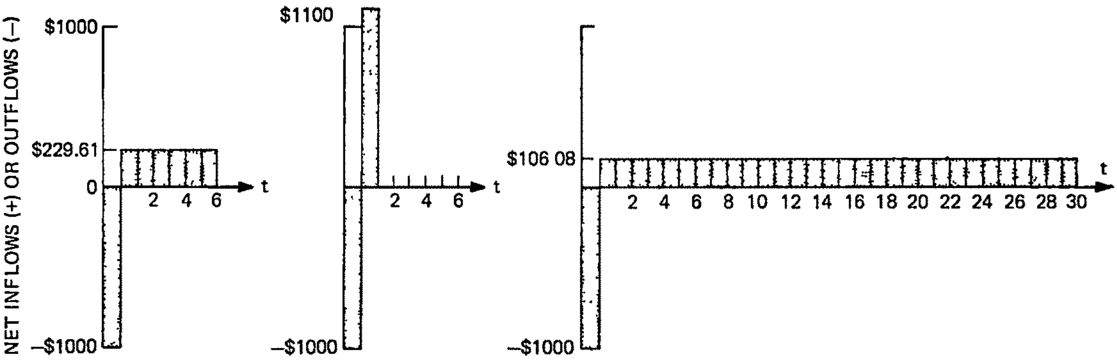
FIGURE 2



But, like expensing policy, the depreciation policy employed has a profound effect on the amount of net book capital, because this is measured by the area under the expensing-depreciation function. Various functions are shown in Figure 2.

In theory, there are infinitely many ways to go from point *A* to point *B* without going out of the rectangle on Figure 2. The five ways shown are not meant to imply that these are the most usual forms—rather they represent potential variants in accounting policy. Each leads to a different estimate of net book value. Given our basic numerical example (that $1000 of outlay generates $229.61 of cash flow per year for six years), the level of net income for the company (assuming equal annual investment) is $377.66, but the level of net book value for the company will range from $0 for Policy No. 1 to $6000 for Policy No. 5. Thus the observable book rate of return will

FIGURE 3



(a) r=10.0 b=12.6

(b) r=10.0 b=10.0

(c) r=10.0 b=15.5

vary from a high of ∞ percent from Policy No. 1 to 6.3 percent for Policy No. 5. Together with expensing policy, depreciation policy produces wide variations in the observable book rate of return for any given DCF rate *r* (which in this example has been held fixed at 10.0 percent).

Other factors affecting the relation of *b* to *r*

■ Variations in accounting policies have a powerful effect on the size of rate of return measured in book rate units relative to the size of rate of return measured in DCF units. But they are not the only factors affecting the ($b \sim r$) relationship. Three other influences are:

1. The economic duration of each investment outlay,
2. The time lag between outlays on the one hand and the commencement of net cash inflows on the other,
3. The time pattern of net cash inflows after they commence.

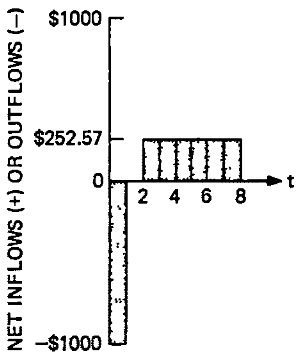
All three of these items can be summarized under a single caption—"The configuration of net cash outflows and inflows" from each investment.

☐ *A*. Figure 3 shows three kinds of assets, all of which yield a DCF rate of 10 percent. Each of these shows outlays (below the horizontal line) and the size and duration of annual inflows (above the line). The first figure, 3a, is a depiction of the standard investment we have used for illustrative purposes thus far: $1000 buys $229.61 a year for 6 years. The DCF rate on this is 10 percent, but the book rate is 12.6 percent.

The second figure, 3b, is a short-duration investment in working capital. $1000 buys an inflow of $1100 in 1 year. The DCF rate (annual compounding basis) is 10 percent. So is the book rate. For this case $b = r$. And possibly this is how someone originated the idea that a rate could be measured accurately by taking the ratio of book income to book value of capital (or possibly it came from another favorite example in elementary commercial arithmetic—the case where $1000 produces a net inflow of $100 a year in perpetuity).

Finally the third figure, 3c, shows a long-duration investment, $1000 outlay producing a net cash inflow of $106.08 a year for 30

FIGURE 4



r=10.0
b=14.7

years. The DCF rate is still 10 percent. But the book rate for such a project, or for a "steady-state" company holding different vintages of such projects, would be 15.5 percent.

In short, in the steady state the longer the "duration" of each asset, the greater the discrepancy in the book rate unit measure relative to the DCF rate unit measure. Since, on working capital alone, the book rate is equal to the DCF rate, otherwise similar companies (with identical DCF rates) would show different book rates merely because each uses a different fraction of working capital relative to depreciable fixed capital.

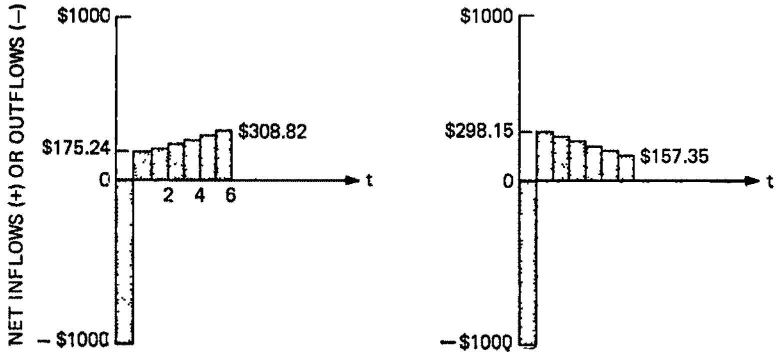
☐ *B.* Figure 4 illustrates the effect of time lags. Beginning with the original $1000 outlay which produces $229.61 a year for 6 years starting one year after the outlay is made, (which has a DCF rate of 10 percent and a book rate of 12.6 percent), this is altered in Figure 4 both to introduce a further one-year lag and to retain, in spite of this, a DCF rate of 10 percent per annum. The annual net cash inflow now has to be larger, $252.57 a year, to compensate for the one-year lag in initial receipts. In book rate units the rate of return will rise from 12.6 percent to 14.7 percent.

In the real world, different assets involve different lags. A truck can be productive the minute it is bought. A hydro-electric plant may take years to build, and many manufacturing plants take time to become fully productive even after they are built. The book rate measure will vary because of these differences.

☐ *C.* Figure 5 illustrates the effect of the time pattern of cash inflows. Even in the absence of inflation, all net inflows do not follow a level path over the assets' expected life. Some assets, especially those involving new products or processes, require time for full market penetration or de-bugging and are therefore likely to show rising net cash inflows. Others, especially new models of old products—like revised editions of automobiles, or college textbooks, or fashion products, or new detergents, or hula hoops—are likely to show a rapid "decay" pattern.

For any assumed DCF rate, the book rate will vary with the pattern. Thus the level pattern in Figure 3a shows a book rate of

FIGURE 5



(a) r=10.0
b=14.1

(b) r=10.0
b=11.0

12.6 percent as against the DCF rate of 10.0 percent. The rising pattern in Figure 5a, which starts at $175.24 per annum and rises at 12 percent a year to $308.82 in year 6, also has a DCF rate of 10 percent, but the book rate is 14.1 percent. Finally the "decay" pattern shown in Figure 5b starts at $298.15 in year 1 and "decays" at 12 percent per annum to $157.35 in year 6. The DCF yield is still 10 percent, but the book rate is 11 percent.

The effect of growth

■ We have shown that for a company in a steady-state situation (i.e., no growth) its book ratio or book rate of return is not an unbiased measure of the true DCF rate of return it is making. Instead we have:

$$b = f(r, x, d, n, w, l, \text{ and } c)$$

which says that the company's observable book rate is a function of many things,

r the DCF rate it is achieving
x its average expensing policy
d its depreciation policy
n average productive life of assets
w the fraction of working capital to total capital
l the average time lag between the outlay for each asset and the commencement of net cash inflows from its use and
c the time pattern of cash inflows.

In this section we lift the steady-state assumption to deal with another major factor which influences the level of the book rate for any given set of the variables outlined above: This is the pace at which the company invests over time. To keep the analysis straightforward we assume that these outlays increase (or decrease) steadily at a given rate g.[8]

If all of the assets it acquires generate a common DCF rate of return equal to r percent, then clearly a company will be earning a DCF rate of r percent regardless of its growth rate g.

But unless the book rate b for such a company is equal to r in the steady state, variations in the pace of growth will cause changes in the observable book rate. The reason for this is simple and can be illustrated clearly in terms of our original "assumed" asset: namely the $1000 outlay which generates a level net cash flow of $229.61 for six years. Regardless of growth rate, a company holding only this form of asset holds six vintages of capital. In the steady state, the book rate earned on each vintage is a function of (a) the expensing and depreciation policy used and (b) the age of the vintage. With full capitalization and a straight-line depreciation policy, the net cash flow and net income attributable to each vintage is constant—at $229.61 *less* $166.66, or $62.95.

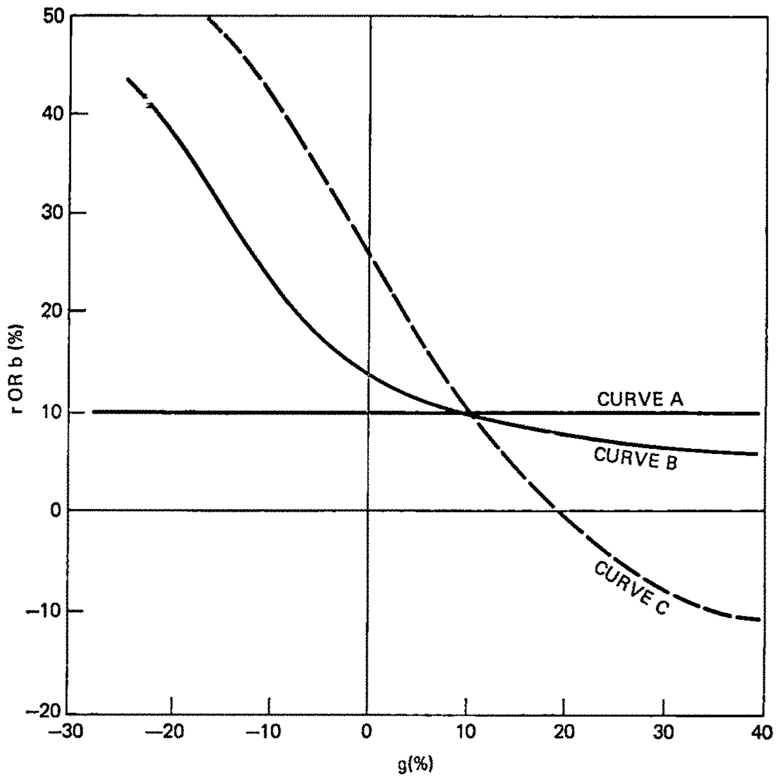
But the midyear net book value of each vintage shrinks with age from [$1000–83.33] for the latest vintage to [$1000–3(83.33)] for the vintage acquired the preceding year, down to $83.33 for the oldest producing asset.

Hence the book rates of return, by vintage, vary: Very low for the mostly undepreciated asset to very high for the almost fully de-

---

[8] Growth here is defined as real growth; i.e., inflation is still assumed to be zero.

FIGURE 6



preciated asset. The book rate for the company (earlier shown to be 12.6 percent) is simply a weighted average of the individual vintage book rates (with the net book values used as weights).

Introducing positive or negative growth into the analysis leaves the individual vintage book rates unaltered, but it changes the relative weights and, hence, the overall company rate (which is its weighted average). With positive growth, the "low yielding" newer vintages get a higher weight relative to the "high yielding" older vintages. Thus the company's overall book rate falls.

In contrast, when growth is negative (i.e., the company's new investment outlays shrink steadily over time), the opposite phenomenon occurs and the company's overall book rate rises.

The relationship of the book rate to the growth rate is shown in Figure 6. Curve A shows the company's DCF rate, i.e., 10 percent, which is of course invariant to the growth rate (since each asset held is earning 10 percent). Curve B shows what the company's book rate would be at various assumed rates of growth (positive and negative). As growth rates become very large the company's overall rate will be dominated by the newest "low yielding" vintage, i.e., curve *B* approaches the yield for the newest vintage. At very fast rates of decay (negative growth) curve *B* approaches the yield for the oldest vintage.

Curve *C* on the same chart shows the same company, but under the assumption that it expenses 50 percent of each investment outlay when the outlay is made and uses a straight-line depreciation policy for the remaining 50 percent. The individual vintage book rates will now be even further apart. The newest vintage, which has charges of

$500 of its acquistion price as an expense in its first year of life, will show a negative book rate of return, whereas the rate of return for the oldest vintage will now be more than twice as high as it was in the 100 percent capitalized situation. The effect of growth on curve C is therefore even more pronounced.

An interesting point on the growth axis is that at which the book rate b is exactly equal to the DCF rate r. This is the same for both curve B and curve C. And it will also hold for other curves D, $E \cdots Z$, which can be drawn for other combinations of the (x, d, n, w, l, and c) variables which affect the book rate of return. Why should this be so? A rigorous proof is beyond the scope of this paper, but one based on intuition runs as follows.

Taking very small intervals of time, we have by definition:

$$b_t = \frac{F_t - D_t}{B_t} \tag{3}$$

This says the book rate is equal to income (cash flow in period t less depreciation and expenses outlays in period t) divided by net book value. Also

$$g_t = \frac{I_t - D_t}{B_t}. \tag{4}$$

This says that the growth rate of investment (which is also the growth rate of net book value) is equal to the net additions to net book value (new investment outlay in period t less depreciation and expense outlays in period t) divided by the existing net book value.

Now the condition $g = r$ (the growth rate equal to the DCF rate) exists only when all of net cash flow generated is reinvested. (An easy way to see this is to think in terms of a barrel of wine increasing in value with age. The annual DCF rate of return on holding this barrel of wine is also the rate at which the barrel increases in value each year—but only if one does not drink part of the contents!) In short $b = r$ only if all the net cash flow is reinvested, i.e. only if $I_t = F_t$.

But if $I_t = F_t$ we can see from equations (3) and (4) above that the conditions $g = b$ must hold. Hence: If

$$g = r$$

we have

$$g = b$$

and thus

$$b = r.$$

In other words, the book rate is an unbiased and accurate measure of the DCF rate for a company which is growing steadily at a rate equal to r (or b).[9] This remarkable equality holds for all values of the extraneous variables x, d, n, w, l, and c.

Mixed assets, price-level changes and irregular growth

■ The introduction of steady real growth moves our explanatory model one step toward reality. The remaining steps would require the introduction and analysis of three other questions.

---

[9] The exact condition is that the company has been growing at this rate for at least n years, where n is the length of each underlying asset's productive life.

1. What is the effect on the $b \sim r$ relationship of mixtures of different assets (and associated variables) within a single company?
2. What is the effect of regular and irregular changes in the price-level of inputs and outputs?
3. Finally, what is the effect of cyclical or other irregular growth rates in investment outlays?

The algebra can be stretched to incorporate some of these real-world variables, at least in simplified form, but the exercise gets increasingly tedious at a faster rate than the rewards in understanding grow. It is easier to jump directly to a fully simulated computerized model in order to explore the relationship of b to r under any assumed set of all these variables. But this is a task for a whole new paper. For the present, it is sufficient to suggest that the points we started out to make have been made:

1. The rate of return in conventional book rate units is conceptually and numerically different from the rate of return in DCF units.
2. Two companies with similar DCF rates of return may well show widely differing book rates of return.

Cost of capital counterparts to *b* and *r*

■ So far we have dealt with rates of return as measures of performance of real assets and the two basically different units in which these rates can be, and are, counted. The same holds true for the other facet: Rate of return as a measure of the "required," "fair," or "target" standard, also known as the "cost of capital." This facet too can be, and is, measured in terms of the two units of account, namely in DCF units or book rate units.

For example, the concept of the embedded cost of debt capital so widely used in utility regulation is a book rate measure. Likewise the "comparative earnings" approach to setting "fair rate of return" which uses book rate information as its basis, is also a book rate measure.

In contrast, there are other methodologies for "measuring the cost of capital" or setting financial standards for capital usage which are clearly DCF type rates. The most obvious example of such a rate is the yield to maturity of actual outstanding or comparably risky bonds. Another is the current dividend yield plus dividend growth rate formula for calculating the cost of equity funds. A less obvious member of the DCF family is the earnings/stock price ratio (commonly called the E/p ratio). This "looks" like a book rate but only because it is measured as a simple ratio. However, close examination shows that it is a short-cut (and frequently unreliable) procedure for measuring a DCF cost of capital rate. The argument in support of this conclusion is as follows. (a) The cost or price of equity capital in any capital market is the DCF rate at which investors discount (or capitalize) future expected benefits from owning equity securities in order to set a current market price for these securities. (b) For non-level projections of benefit streams, the investors' discount rate can be found only by a trial and error process. For dividend streams that grow continuously at a constant rate the mathematics of the process reduces to a fairly simple form, generally expressed as

$$k = \frac{D_0}{P_0} + g$$

where k is the DCF discount rate being solved for, D_0 is the current rate of dividends per share, P_0 is the current price per share, and g is the rate at which dividends are expected to grow.[10]

With even more restrictive assumptions about future investor expectations, the mathematics of finding the discount rate can be simplified even further. Thus, if future earnings (E) from existing assets can be assumed to be a level, perpetual stream, the present value equation linking current price to future benefits simplifies to[11]

$$k = \frac{E}{p}.$$

In short (quite apart from the validity of the almost impossibly restrictive assumptions underlying this formulation), the E/p basis for estimating k is simply a reduced form of the equation used in solving for a DCF rate.

Thus on the "financial standard" side we are faced with the same kind of confusion as we have on the "performance" side: There are two distinct and numerically different units, one a DCF unit usually symbolized as k, and one a book rate unit, which I will here symbolize as β. Both are called "cost of capital," and both use the label "percent per annum required rate of return."

Uses and misuses of disparate "rates"

■ The rules for the "correct" usage of these concepts and measures are fairly simple. If the actual or prospective performance of any investment is measured in DCF units, and if this rate is being assayed against some target or "reasonable rate standard," it is clear that the relevant standard must itself be calculated in DCF units. In short, r must be matched against k in order to produce rational decisions and judgments. Sometimes adjustments may be required, either to r or to k to allow for perceived differences in riskiness between the kind of investment being assayed and the kind of investments from which the k measure has been derived.

By the same token, if book rate units, b, are used to measure actual or prospective performance, the proper standard of comparison is against β. Here too risk adjustments are legitimate. In addition, adjustments may be necessary if the investment or collection of investments being assayed differ significantly with respect to the set of variables (x,d,n,w,l,c, and g) from the collection of investments from whose performance the estimate of β has been derived.

The potential misuses of these tools of thought involve all the inconsistent comparisons which can be made among the four measures. What is surprising is that almost every conceivable form of misuse is being practiced today. Some examples are:

1. A regulatory authority measures the cost of capital in DCF units (k) and then translates this number into "required revenues" by multiplying it against a net book value estimate (or one based on net book value).
2. The same authority measures the cost of equity capital (k_e) in DCF units and the cost of debt in book rate units (embedded cost), and uses the weighted average of these two rates as the figure to be multiplied against a net book value rate base.

---

[10] For a proof, see [5].

[11] For a proof, see [15], p. 25, note 7.

But regulation is not the only arena in which the disparate units are used as if they were interchangeable estimates of a common concept. Companies which have moved toward the measurement of investment worth in terms of a promised DCF rate still use book rates as a basis for setting the financial standard rate against which this promised DCF can be compared.

In some cases, DCF rates are used for *ex-ante* capital budgeting purposes, but an unadjusted book rate measure is used for later (*ex-post*) audit purposes, to check whether or not the investment lived up to its promise.

Finally, there is still a great deal of implicit acceptance that widely differing observed book rates are unbiased measures of actual profitability.

Understanding and avoiding these potential misuses of "rates of return" will not in itself provide correct answers. Important differential effects of price level changes on all measures, and the random effects of estimating errors, remain as significant barriers to be overcome, both for regulators and for private managers. Meanwhile, understanding that book rate measures and DCF rate measures are not different estimates of the *same* thing but rather estimates of *different* things should eliminate at least part of the confusion surrounding "rates of return on investment."

References

1. Bierman, H. "Depreciable Assets—Timing of Expense Recognition," *The Accounting Review*, October 1961, pp. 613–18.
2. Carlson, R. *Measuring Period Profitability.* Unpublished doctoral dissertation, Stanford University, 1967.
3 Cottle, S. and Whitman, W. *Corporate Earning Power and Market Valuation.* Durham, N. C.: Duke University Press, 1959.
4 Gordon, M. J. "The Payoff Period and the Rate of Profit," *The Management of Corporate Capital*, edited by Ezra Solomon. New York: MacMillan-Free Press, 1959, pp. 48–55.
5. ——— and Shapiro, E. "Capital Equipment Aanlysis: The Required Rate of Profit," *The Management of Corporate Capital*, edited by Ezra Solomon. New York MacMillan-Free Press, 1959, pp. 141–47.
6. Harcourt G. C. "The Accountant in a Golden Age," *Oxford Economic Papers*, New Series, Vol. 17, No. 1 (1965).
7. Johnson, O. "Two General Concepts of Depreciation," *Journal of Accounting Research*, Spring 1968, pp. 29–37 .
8. Kuznets, S., ed. *Income and Wealth, Series II.* Cambridge: International Association for Research into Income and Wealth, Bowes-Bowes, 1952.
9 Lanzillot, R. F. "Pricing Objectives in Large Companies," *American Economic Review*, Vol. 48 (1968), pp. 921–40.
10 Laya, J. C. *A Cash-Flow Model for Rate of Return.* Unpublished doctoral dissertation, Stanford University, 1965.
11. Minhas, B S. *An International Comparison of Factor Costs and Factor Use.* Amsterdam: North Holland Publishing Company, 1963.
12. Solomon, E., ed. *The Management of Corporate Capital.* New York: MacMillan, 1959.
13. ———. "Research on Return on Investment: The Relation of Book-Yield to True Yield," in American Accounting Association, *Research in Accounting Measurement*, 1965.
14. ———. "Systematic Errors in Book Rates of Return," Society of Petroleum Engineers, SPE 655 (1963)
15. ———. *The Theory of Financial Management.* New York: Columbia University Press, 1963.

16. ———. "The Variation Between True Yield and Book Rate of Return in the Oil and Gas Producing Industry," Testimony before Federal Power Commission, Area Rate Proceedings AR-61-1 (1961).
17. ——— and LAYA, J. C. "Measurement of Company Profitability: Some Systematic Errors in the Accounting Rate of Return," in *Financial Research and its Implications for Management Decisions*, A. A. Robichek, ed. New York: John Wiley & Sons, 1966, pp. 152–183.
18. STIGLER, G. J. *Capital and Rates of Return in Manufacturing Industry*. Princeton, N. J.: Princeton University Press, 1963.
19. VATTER, W. J. "Income Models, Book Yield and the Rate of Return," *The Accounting Review*, October 1966, pp. 681–98.
20. WRIGHT, F. W. "Toward a General Theory of Depreciation," *Journal of Accounting Research*, Spring 1964, pp. 80–90.

Some analytical results on tax depreciation

Peter B. Linhart
Manager—Financial and Regulatory Analysis
American Telephone & Telegraph Co.

A mathematical framework for the discussion of accelerated tax depreciation methods for regulated utilities is constructed. Using this framework a number of results are derived, some of them novel. We show, for example, under what circumstances interested parties such as the customers and the tax collector should rationally prefer flowthrough or normalization. The most important conclusion is that, following a switch from straight-line book and tax depreciation to accelerated tax depreciation with flowthrough, very substantial rate increases will always be required after a fixed interval, regardless of the firm's growth rate. Effects of accelerated tax depreciation on interest coverage and tax-free dividends are also discussed, and some numerical results are given. The model is general enough to allow incorporation of a number of realistic conditions, as, for example, debt in the capital structure and a dispersed life table.

1. Introduction

■ In this paper, we construct an analytical model within which various aspects of accelerated tax depreciation can be discussed—most importantly, the question of the effects of various tax depreciation policies on utility customers' bills. These issues have been the subject of argument and confusion since 1954, and most recently in the discussion of the new tax law passed by Congress in January of this year.

In 1954, Congress authorized the use of accelerated tax depreciation [1]. The intent of this legislation was to "assist modernization and expansion of industrial capacity." This would occur because tax deferral would "increase available working capital." (Note that tax deferral, not tax reduction, is involved; the total amount of tax depreciation deductions is not increased.) Two options were thus opened to nonregulated companies: 1) they can use accelerated depreciation for both tax purposes and for accounting purposes (and reporting their earnings), or 2) they can stick to straight-line depreciation for accounting purposes, while taking accelerated tax depreciation. If option 2 is selected, both the SEC [2] and the accounting profession [3] require that a reserve (the "normalization reserve")

The author is Manager—Financial and Regulatory Analysis in the Management Sciences Division of the American Telephone and Telegraph Company. He holds the B.A. and M.A. degrees in physics and received the Ph D degree in applied mathematics from Columbia University in 1963. He has taught physics at the University of California (Berkeley) and at Columbia University, was a Research Associate at IBM-Watson Laboratories from 1954 to 1956, and joined Bell Telephone Laboratories in 1956. Dr. Linhart came to the AT&T Company in 1967, where he is presently working on a financial model of the Bell System.

The work presented in this paper was done under the direction of H. J. McMains of AT&T. The author is indebted to many colleagues and friends for ideas and technical help—most particularly, Mr. A. L. Stott of AT&T, Professor Harold Bierman, Jr., of Cornell, and Dr. F. W. Sinden of Bell Telephone Laboratories.

be established to reflect the reduction in current tax payments: that is, they require that the premature using up of part of the value of an asset of the company (its right to recover its capital free of tax) be recognized as a current cost, not an increase in earnings.

The situation for regulated utilities is different. Regulatory authorities generally require straight-line depreciation for accounting and rate-making purposes and for reporting to stockholders [4]. Thus a utility which wishes to use accelerated tax depreciation must keep two sets of books. Moreover, many regulatory authorities prohibit the establishment of a normalization reserve. They insist that the reduction in current tax payments (from the accelerated use of tax depreciation deductions) be treated as an increase in earnings, making it possible to reduce rates, and thus "flow through" the effect of the tax deferrals to present customers. (The terms "normalization" and "flowthrough" will be defined precisely in the mathematical development which follows.) In more than one case [5] a regulatory body has imputed flowthrough, i.e., set rates as if flowthrough accounting had been used, to a utility using straight-line depreciation for both book and tax purposes. In defense of this point of view it is argued that the tax postponement associated with accelerated tax depreciation is in effect a tax reduction, since the deferred taxes for a growing company will always be more than offset by tax reductions on new plant. Thus we have a conflict between regulatory bodies urging flowthrough and certain utilities (for example, AT&T) that would like to use accelerated tax depreciation as a source of working capital through "normalization."

Weighty debate on tax depreciation has occurred, for example, in the FCC investigation of AT&T (Docket 16258) and before the House Ways and Means Committee [6] and the Senate Finance Committee. However, in the absence of an analytical framework the arguments have tended to be qualitative, or moralistic, or to rely on particular numerical estimates and examples.

The purpose of this paper is to fill this gap by providing an analytical framework within which the effects of various depreciation policies on utilities, their customers, and the Federal treasury can be examined with some precision and generality. Although numerical examples are given, the emphasis is theoretical. The main application we have in mind is to a regulated utility. Therefore we will assume a return on investment which is a fixed fraction of net plant, and that straight-line depreciation must be used for book or rate-making purposes, although accelerated tax depreciation may be used.

Since we are interested in insight rather than numerical results, we want to use a model which is as simple as it can be while preserving the essential features of the problem. Thus, while we include in our description such features as a dispersed life table, debt, and growth of the firm, we exclude, for example, consideration of changing depreciation rates, the variety of categories of plant, and salvage. However, the techniques described can easily be elaborated by those interested in "exact" quantitative answers. A precise description of the model appears in section 2.

We do not treat in the present paper a variety of important and difficult economic aspects of the depreciation problem—e.g., questions of technological obsolescence, elasticity of demand, inflation. Neither do we enter (explicitly) the debate about "best" depreciation policies, our aim being to provide tools rather than answers. One aspect of the problem which we consider important is what might be called the informal game-theoretic aspect. That is, a decision on depreciation policy must be looked at in terms of its effects on the various "players," who are the stockholders, the bondholders, the tax-collector, and the firm (if such an entity can indeed be meaningfully defined). We shall see that the interests of the various players are partly coincident and partly opposed.

The reader who wishes to skip at this point to a summary of the principal results will find them in Section 6.

2. Description of the model, notation, and some elementary properties

■ It is convenient to begin by considering a single vintage of plant: that is, all the new plant purchased and installed by the firm in a given year. We assume that this plant is all installed at the beginning of the year, say at $t = 0$. We adopt the convention that $t = 0$ is the beginning of year 1, and in general $t = j - 1$ is the beginning of year j.

There are certain capital costs associated with this plant in each year of its life, e.g., in year j. These are:

(1) return = ρX_j, where ρ is the allowed rate of return and X_j is the net book plant (= rate-base = plant not yet depreciated for book purposes) at the start of year j. (Note the implicit assumption at this point that the company succeeds in earning its allowed return, i.e., that its *actual* rate of return = ρ.)[1]
(2) taxes = T_j; these are simply the actual tax dollars paid to a single agency, say the Federal Government, in year j.
(3) book depreciation = D_{B_j}.

The total book depreciation over the life of the vintage must equal the initial investment. That is, if the firm paid K dollars for this vintage, then

$$\sum_{j=1}^{\infty} D_{B_j} = K. \tag{1}$$

The upper limit of ∞ on the sum in equation (1) is meant to suggest that any vintage may have a few very long-lived survivors.

In the case of "normalization," there is another component of capital cost, namely a contribution to a reserve for deferred taxes. We neglect this at present and describe it in detail later.

All these capital costs must be collected as revenues from the firm's customers. Thus the revenue requirements in year j are[2]

$$RR_j = \rho X_j + T_j + D_{B_j}. \tag{2}$$

---

[1] A glossary of symbols appears in Appendix 2.

[2] It would be a mistake to suppose, even in the case that the firm succeeds in earning its allowed rate of return, that a single item or even a single vintage of plant generates in each year revenues equal to RR. In practice many items, and perhaps even many vintages, working in conjunction provide the service that produces the revenue. Moreover, even to the extent that revenue generation can be attributed to a single item, revenues will be approximately level over the life of the item. But RR declines over the life of the item since it contains the term ρX. Nevertheless, our model, when applied to a superposition of vintages, does produce correct revenue requirements, and if the firm just succeeds in earning its allowed rate of return, it does earn revenues, R, equal to revenue requirements, RR. The reconciliation of these two points of view, through the calculation of "levelized" revenue requirements attributed to a single item of plant, is a problem in engineering economics not treated in this paper.

(We reserve the notation R for revenues in cases when these do not equal revenue requirements.)

In reality, the firm must collect from its customers revenues to cover not only capital costs but all other expenses. However, by RR we mean revenue requirements net of these other expenses.

Now both tax depreciation and interest expense are tax deductible. Let

D_{T_j} = tax depreciation in year j,
i = imbedded cost of debt, and
δ = debt-ratio (ratio of debt to total capital).

It is assumed that debt is paid off in the same time-pattern that the asset is depreciated, so that in year j the interest expense is $i\delta X_j$.

Thus, the taxes paid in year j are

$$T_j = T_0(RR_j - D_{T_j} - i\delta X_j)\,, \tag{3}$$

where T_0 is the Federal Income Tax rate.

We shall assume that the tax basis is the same as the book basis, i.e.,

$$\sum_{j=1}^{\infty} D_{T_j} = K\,. \tag{4}$$

In what follows, the year-subscript, j, will be omitted wherever the context precludes confusion.

Equations (2) and (3) may be regarded as two simultaneous equations in the two unknowns RR and T. Their solutions are

$$RR = \frac{1}{1 - T_0}[(\rho - T_0 i\delta)X + D_B - T_0 D_T]\,, \tag{5}$$

and

$$T = \frac{T_0}{1 - T_0}[(\rho - i\delta)X + D_B - D_T]\,. \tag{6}$$

The quantity $(\rho - T_0 i\delta)$ occurring in equation (5) is of some interest and will occur repeatedly below. It is the cost of capital to the firm, taking account of the tax-deductibility of interest.

When the firm uses the same depreciation method for both book and tax, we have $D_B = D_T$, and equations (5) and (6) simplify to

$$RR = \frac{1}{1 - T_0}(\rho - T_0 i\delta)X + D_B\,, \tag{5a}$$

and

$$T = \frac{T_0}{1 - T_0}(\rho - i\delta)X\,. \tag{6a}$$

Equations (5a) and (6a) apply, for example, in the important case in which straight-line depreciation is used for both book and tax purposes.

☐ **Flowthrough vs. Normalization.** Taking accelerated depreciation for tax but not for book purposes means that $D_T > D_B$ in the early years of the life of an asset, and $D_T < D_B$ in the later years. That is, from equation (3), taxes are lower in the earlier years and higher in the later years than if straight-line depreciation were used for both book and tax purposes (or indeed any other identical method), and, from equation (2), so are revenue requirements. However, the total taxes over the life of the vintage are independent of the tax depreciation method, as may be seen by summing equation (6) over the years of the life and substituting equation (4) in the result. (The same is true of revenue requirements.) In other words, when accelerated depreciation is taken for tax purposes, taxes are deferred. If no allowance for a reserve for deferred taxes is included in the revenue requirements, the resulting phenomenon is known as flowthrough. Flowthrough clearly entails a risk to the company, the risk that the company will be unable in the later years to collect enough revenues from its customers to pay its deferred taxes without a reduction in its actual rate of return. In order to collect such revenues, two conditions must be satisfied:

(1) the relevant regulatory commissions must allow rate adjustments such that a rate of return ρ can be maintained in the later years of the life, and

(2) the elasticity of the market must be such that it is possible, by means of rate adjustments, to recover the required revenues in the later years.

There are those who believe that this increment of risk associated with flowthrough should be compensated for by an increase in the allowed rate of return, ρ, although the appropriate amount of such an increase is difficult to determine. On the other hand, there are those who argue that the risk of flowthrough is without force, since as long as the firm is growing (or even not shrinking) the higher taxes on the older vintages are overbalanced by the lower taxes on the larger newer vintages. This latter argument is incorrect, as we shall prove in a later section (see Theorem 5 on page 100).

Under *normalization*, by contrast, an allowance for an increment to a reserve for deferred taxes (the "normalization reserve") is included in the revenue requirements. This allowance[3] is

$$T_0(D_T - D_B),$$

which represents the direct effect in reducing taxes of accelerated depreciation, but not the "doubling" effect, which is included in the term

$$\frac{T_0}{1 - T_0}(D_T - D_B)$$

in equation (6).

That is, with normalization, we modify equation (2) to read

$$RR_j = \rho Y_j + T_j + D_{B_j} + T_0(D_{T_j} - D_{B_j}). \qquad \text{(2b)}$$

The new symbol, Y, requires some explanation. We assume that, since the normalization reserve represents in some sense a supply of interest-free capital, the firm will not be permitted, or will not elect, to earn on this amount. Consequently it earns at a rate ρ on a reduced rate-base, which at the start of year j is

---

[3] Part of the value of an asset is its tax-reducing power. Accelerated tax depreciation uses up this value faster, and thus represents an additional expense. Under normalization this expense is recognized by inclusion of the quantity $T_0(D_T - D_B)$ in the revenue requirements.

$$Y_j = X_j - N_j \,, \tag{7}$$

where N_j is the normalization reserve at the start of year j:

$$N_j = \sum_{k=1}^{j-1} T_0 \, (D_{T_k} - D_{B_k}) \,. \tag{8}$$

At the start of the life of the vintage in question, $N = 0$, and at the end of the life, by substitution of equations (1) and (4) in equation (8), N is also zero.

Equation (3) is unchanged under normalization, and the solutions of the simultaneous equations (2b) and (3) are

$$RR = \frac{1}{1 - T_0} (\rho - T_0 i\delta) Y + D_B \,, \tag{5b}$$

$$T = \frac{T_0}{1 - T_0} (\rho - i\delta) Y + T_0(D_B - D_T) \,. \tag{6b}$$

A couple of points are worth noting. First, equation (5b) resembles equation (5a); that is, the pattern of revenue requirements is the same under normalization as when the same depreciation method (e.g., straight-line) is used for both book and tax, except for the decreased rate-base. Since $N > 0$, $Y < X$. Hence, comparing (5b) with (5a), we see that the total revenue requirements over the life of the vintage are less under normalization than under flowthrough or when straight-line depreciation is used for both book and tax purposes. From (6b) the total taxes are also less. Second, the total tax expense, by which we mean the sum of the taxes actually paid and the amount added to the normalization reserve, is in any year given by

$$\frac{T_0}{1 - T_0} (\rho - i\delta) Y \,.$$

Thus under normalization, as when straight-line depreciation is used for both book and tax purposes, the total tax expense is always proportional to the return:

return ÷ total tax expense

$$= \rho Y \div \frac{T_0}{1 - T_0} (\rho - i\delta) Y = \rho(1 - T_0) \div T_0(\rho - i\delta) \tag{9}$$

independently of time.

Under flowthrough, this proportionality fails to hold. In fact, the violation of this proportionality could be taken as a generalized definition of flowthrough.

☐ **The Life Table.** By the life table, we will mean the fraction of a vintage (measured in dollars) surviving at age t. We denote this function by $f(t)$. Of course,

$$f(0) = 1 .$$

We assume $f(t)$ to be the same function for every vintage.

The function $f(t)$ is the complement of a distribution function. That is, if the random variable χ represents the life of an item of plant, then

$$g(t) \equiv 1 - f(t) = Pr\{\chi < t\}$$

is a distribution function. Now the mean, if it exists, of a non-negative random variable with distribution function g is given by either of the expressions

$$\int_{t=0}^{\infty} t dg(t) = \int_{t=0}^{\infty} [1 - g(t)] dt .$$

Therefore, if the mean life of a vintage is L, we have

$$L = \int_{t=0}^{\infty} f(t) dt . \qquad (10)^4$$

That is, the mean life is the area under the life table.

The simplest life table is a square life table, defined by

$$f(t) = \begin{cases} 1, & t \leq L , \\ 0, & t > L . \end{cases} \qquad (11)$$

This means that for a square life table the entire vintage drops dead at age L.

Very different from a square life table is an exponential one:

$$f(t) = e^{-t/L} . \qquad (12)$$

With a square life table, the variance of the life is 0. With an exponential life table, the variance is L^2. Most cases of interest lie somewhere in between.

☐ **Several Depreciation Schedules.** The depreciation methods we shall be concerned with are straight-line (henceforth abbreviated *SL*), double declining balance (*DDB*) and sum of the year's digits (*SYD*). It is easiest to define these first with respect to a vintage of plant with a square-life table and mean life L.

For *SL* depreciation, the depreciation rate is $1/L$ for every year of the life; i.e., if the initial investment is K, the depreciation in the j'th year of the life is

$$D_j = \begin{cases} K/L, & (j = 1, 2, \ldots , L) , \\ 0, & (j > L) . \end{cases} \qquad (13)$$

For *DDB*, the depreciation rate is $2/L$, but is applied to the *net* (undepreciated) plant, rather than to the *gross* plant (survivors).

---

[4] In discussing dispersed life tables, it is useful to think of time as a continuous variable. In general, we shall use the discrete or continuous representation (sums or integrals), whichever is more convenient. Discrete (annual) notation is more like accounting practice, while continuous is more in accord with the real patterns of plant construction, etc.

Thus, the depreciation in the j'th year of the life is

$$D_j = \frac{2}{L} X_j, \ (j = 1, 2, \ldots). \tag{14}$$

Since $X_1 = K$ and $X_j = X_{j-1} - D_{j-1}$, equation (14) determines a recursion whose solution is

$$D_j = \frac{2K}{L}\left(1 - \frac{2}{L}\right)^{j-1}. \tag{15}$$

For SYD, one takes L parts of the total depreciation in year 1, $L - 1$, parts in year 2, ..., 1 part in year L. That is,

$$D_j = \begin{cases} \dfrac{2K(L + 1 - j)}{L(L + 1)}, & (j = 1, 2, \ldots, L), \\ 0, & (j > L). \end{cases} \tag{16}$$

The continuous equivalents of the above equations may be of some interest. Here we redefine $D(t)$ as the depreciation *rate* at time t; that is,

$$D(t) = \lim_{\Delta \to 0} \frac{\{X(t + \Delta) - X(t)\}}{\Delta}.$$

We then have:

for SL,

$$D(t) = \begin{cases} \dfrac{K}{L}, & 0 \leq t \leq L, \\ 0, & t > L, \end{cases} \tag{13a}$$

for DDB,

$$D(t) = \frac{2K}{L} e^{-2t/L}, \quad 0 \leq t, \tag{15a}$$

for SYD,

$$D(t) = \begin{cases} \dfrac{2K(L - t)}{L^2}, & 0 \leq t \leq L, \\ 0, & t > L. \end{cases} \tag{16a}$$

Note that strictly DDB depreciation continues after the physical death of the vintage (see, however, Appendix 1, near equation (A8)).

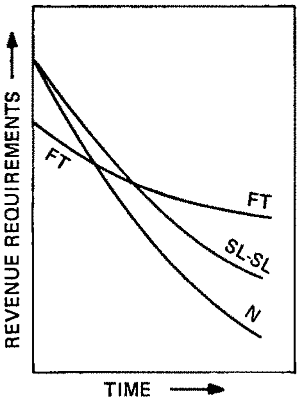
For a dispersed life table, all of the above becomes considerably more complicated. We must now distinguish between two methods of applying the depreciation method to the life table, namely the *ELG* (equal-life-group) method, and the *VG* (vintage-group) method. Details are given in Appendix 1.

3. Results on present worth

■ We have seen that for a single vintage of plant with an arbitrary life table, the stream of revenue requirements over the life of the vintage has the following characteristics:

(1) the total revenue requirements over the life are the same for *SL-SL* (straight-line for both book and tax) as for *FT* (flowthrough);
(2) revenue requirements under *FT* start lower and end higher than those under *SL-SL*;
(3) the total revenue requirements under *N* (normalization) are less than under *SL-SL* or *FT*;
(4) revenue requirements under *N* have initially the same value as under *SL-SL*, but are lower for all $t > 0$ (compare equations (5a) and (5b)).

FIGURE 1



Thus the revenue requirements have the qualitative behavior shown in Figure 1.[5]

Now it is the customer who must supply these revenues. Which depreciation method should he prefer? (Of course, he must really pay the revenue requirements on a multiplicity of vintages; we extend our results to this case later.) The customer presumably prefers small bills to large ones; hence he prefers *N* to *SL-SL*, since the latter dominates the former. But since the *FT* curve crosses both the *N* and *SL-SL* curves, it is not clear what his (rational) preferences should be as between *FT* and *SL-SL* or between *FT* and *N*. In a case like this one may resort to some one-parameter criterion for comparing time series of cash flows. We use the common one: present worth. This is defined as follows: the present worth of a sequence $a_1, a_2, \ldots,$ at a (positive) discount-rate r, referred to the start of period 1, is

$$PW_r(a) = \sum_{j=1}^{\infty} a_j(1 + r)^{-j}, \qquad (17)$$

if the series converges.[6]

☐ **The Customer's Point of View.** In the present case, the customer prefers the depreciation method which minimizes $PW_r(RR)$. We shall see that the customer's preference depends on r, but we do not go into the difficult question of determining the numerical value of r (i.e., the time-value of money to customers).

First, we give a Lemma, which will be useful.

LEMMA 1 (*present worth of a sequence of partial sums*): *Consider an infinite sequence*[7] *of numbers*

$$(a) = a_1, a_2, \ldots,$$

and also the sequence of partial sums

$$(b) = b_1, b_2, \ldots,$$

where

$$b_j = \sum_{k=1}^{j-1} a_k, \quad (b_1 = 0).$$

---

[5] For very short mean lives the *FT* revenue requirements actually increase during the life, but this does not change the argument.

[6] The convergence criterion is this: there must exist a J such that for all $j > J$, $1 + r > |a_{j+1}/a_j|$.

[7] Finite sequences are a special case, obtained by setting all the a's after a certain one equal to zero.

Then

$$PW_r(b) = \frac{1}{r} PW_r(a). \tag{18}$$[8]

We consider next the comparison between flowthrough and normalization. Let the customer's saving from N, compared to FT, be, in year j,

$$\Delta RR_j \equiv RR_{FT_j} - RR_{N_j}.$$

From equations (5) and (5b),

$$\Delta RR_j = \frac{1}{1 - T_0} [(\rho - T_0 i\delta)N_j + T_0(D_{B_j} - D_{T_j})], \tag{19}$$

where N_j is defined by equation (8).

Taking the present worth of equation (19) at discount rate r, referred to $t = 0$,

$$PW_r(\Delta RR) = \frac{1}{1 - T_0} [(\rho - T_0 i\delta) PW_r(N) + T_0 PW_r(D_B - D_T)].$$

But by Lemma 1,

$$PW_r(N) = -\frac{T_0}{r} PW_r(D_B - D_T).$$

Therefore,

$$PW_r(\Delta RR) = \frac{T_0}{1 - T_0} \left[1 - \frac{1}{r}(\rho - T_0 i\delta)\right] PW_r(D_B - D_T). \tag{20}$$

It is apparent from equation (20) that PW_r (ΔRR) vanishes[9] at the critical discount rate

$$r_1 = \rho - T_0 i\delta. \tag{21}$$

Note that, from equations (1), (4), and (17),

$$PW_0(D_B) = \sum_{j=1}^{\infty} D_{B_j} = K = \sum_{j=1}^{\infty} D_{T_j} = PW_0(D_T).$$

For any $r > 0$, $PW_r(D_B - D_T) < 0$.
Thus we conclude that:

$$\text{for} \quad 0 < r < r_1, \quad PW_r(\Delta RR) > 0,$$
$$\text{for} \quad r > r_1, \quad PW_r(\Delta RR) < 0.$$

---

[8] The proof of Lemma 1 is as follows.

Let $z = 1/(1 + r) < 1$

$$PW_r(b) = \sum_{j=1}^{\infty} b_j z^j = \sum_{j=1}^{\infty} z^j \sum_{k=1}^{j-1} a_k = \sum_{k=1}^{\infty} a_k \sum_{j=k+1}^{\infty} z^j = \sum_{k=1}^{\infty} a_k \frac{z^{k+1}}{1 - z} = \frac{z}{1 - z} \sum_{k=1}^{\infty} a_k z^k = \frac{z}{1 - z} PW_r(a).$$

But $z/(1 - z) = r$, whence

$$PW_r(b) = \frac{1}{r} PW_r(a), \quad \text{Q.E.D.}$$

[9] This might suggest that there is a second solution to the equation $PW_r(\Delta RR) = 0$, namely $r = 0$. However, this is not the case. When $r = 0$, equation (20) is indeterminate, returning to first principles, $PW_0(RR_{FT}) > PW_0(RR_N)$, since these present worths are just sums (see the discussion following equation (6b)). Therefore, the solution expressed in equation (21) is unique.

We have now proved

THEOREM 1: *For a single vintage, the rational customer should prefer normalization to flowthrough if his personal discount rate is less than* r_1, *be indifferent if it equals* r_1, *and prefer flowthrough to normalization if it is greater than* r_1 (*where* r_1 *is defined by equation* (*21*)).

Economic Interpretation. Theorem 1 is susceptible of a simple economic interpretation. Under *flowthrough,* the firm at first reduces the customer's bills by the full amount of its tax reduction corresponding to accelerated tax depreciation, and later increases his bills by the full amount corresponding to the payment of deferred taxes. Thus the firm in effect lends the customer this amount. Under *normalization* no such loan is made, but the customer's bills are reduced by an amount equal to the reduction in revenue requirements corresponding to the exclusion of the normalization reserve from the rate base. Thus the question reduces to the following: can the customer invest this loan (under flowthrough) so as to earn more or less than the reduction in revenue requirements under normalization?

To be more specific, suppose that, because of accelerated tax depreciation, D_T is increased in year 1 by an amount Δ and decreased in year 2 by the same amount. Then, under flowthrough the customer has for one year the use of an amount of money equal to, by equation (5),

$$\frac{T_0}{1 - T_0}\cdot\Delta .$$

On this, he can earn

$$r\cdot\frac{T_0}{1 - T_0}\cdot\Delta ,$$

where r is his discount-rate.

Under normalization, his bills are reduced in this year by (Equation 5b)

$$\frac{\rho - T_0 i\delta}{1 - T_0}(Y - X) = \frac{\rho - T_0 i\delta}{1 - T_0}\cdot T_0\Delta .$$

Comparison now yields Theorem 1, namely, that the critical value of r is $\rho - T_0 i\delta$.

Further clarification may result from the following: let ρ_e be the rate of return on equity. Then

$$\rho = \rho_e(1 - \delta) + i\delta .$$

But the pre-tax rate of return, which is what the customer sees, is (since interest is tax-deductible)

$$\frac{1}{1 - T_0}[\rho_e(1 - \delta)] + i\delta = \frac{1}{1 - T_0}(\rho - T_0 i\delta) = \frac{1}{1 - T_0}\cdot r_1 .$$

Actually, the numerical differences are small. Consider for example a firm for which

$$\rho = 0.1\,,$$

$$\delta = 0.5\,,$$

$$i = 0.05\,,$$

and assume

$$T_0 = 0.5\,.$$

Then

$$r_1 = 0.0875\,.$$

For a vintage with a square life table, initial value \$1000 and mean life 20 years, Table 1 shows $PW_r(RR_{FT})$ and $PW_r(RR_N)$ for various discount rates.

TABLE 1

r	$PW_r(RR_{FT})$	$PW_r(RR_N)$
0.00	2837.50	2560.42
0.05	1847.45	1776.37
0.08	1489.39	1478.60
0.0875	1417.33	1417.33
0.10	1309.45	1324.61
0.12	1163.39	1196.91
0.15	991.26	1042.61
0.20	788.67	854.35

The comparison between *SL-SL* and *FT* is also simple. Since the total revenue requirements over the life of the vintage are equal in the two cases, and since RR_{FT} starts out lower than RR_{SL},

$$PW_r(RR_{FT}) < PW_r(RR_{SL})\,,$$

and the customer should prefer *FT* for all $r > 0$. For $r = 0$ he should be indifferent.

☐ **The Tax Collector's Point of View.** So much for the customer's point of view. Next consider the tax collector. Let us assume he prefers high tax receipts. Define

$$\Delta T_j = T_{FT_j} - T_{N_j}\,.$$

From equations (6) and (6b),

$$\Delta T_j = \frac{T_0}{1 - T_0}\,[(\rho - i\delta)N_j + T_0(D_{B_j} - D_{T_j})]\,.$$

Again using Lemma 1,

$$PW_r(\Delta T) = \frac{T_0^2}{1 - T_0}\left[1 - \frac{1}{r}(\rho - i\delta)\right]PW_r(D_B - D_T)\,. \qquad (23)$$

From the point of view of the tax collector, the critical discount rate is $r_2 = \rho - i\delta$.

For the numerical example used above, $r_2 = 0.075$. Equation (23) proves

THEOREM 2: *For a single vintage, the rational tax collector should prefer flowthrough to normalization if his discount rate is less than r_2, be indifferent if it equals r_2, and prefer normalization to flowthrough if it is greater than r_2.*[10]

Economic Interpretation. The economic interpretation of Theorem 2 is parallel to that of Theorem 1. If D_T is increased by Δ, as before, in year 1, and the reverse is true in year 2, then under flowthrough, taxes paid are reduced in year 1 by (according to equation (6))

$$\frac{T_0}{1 - T_0}\Delta ,$$

and increased by the same amount in year 2. Under normalization, on the other hand, taxes are reduced in year 1 by (see equation (6b))

$$T_0\Delta$$

and increased in year 2 by the same amount (ignoring for the moment the term in equation (6b) involving Y). Thus, under normalization the tax-collector has the use of

$$\left(\frac{T_0}{1 - T_0} - T_0\right)\Delta = \frac{T_0{}^2\Delta}{1 - T_0}$$

additional dollars, on which he can earn

$$\frac{T_0{}^2\Delta}{1 - T_0}\cdot r ,$$

where r is his discount rate. On the other hand (referring now to the neglected term in equation (6b)), he is denied the taxes associated with N, namely

$$\frac{T_0{}^2}{1 - T_0}(\rho - i\delta)\Delta .$$

Comparison now yields Theorem 2, namely that the tax collector's critical discount rate is $\rho - i\delta$.

Note that $\rho - i\delta = (1 - \delta)\rho_e$. Thus this theorem reflects the fact that the tax-collector cares only about equity earnings, interest being tax-deductible.

☐ **The Firm's Point of View.** If the firm wishes to minimize revenue requirements, its situation is described by Theorem 1, and its interests coincide with those of the customers. Suppose, however, that the firm wishes to maximize the present worth of cash flows[11] at some discount rate, where cash flow is defined by

$$C = R - T$$

---

[10] The form of Theorem 2 is slightly different from that of Theorem 1 because the tax collector prefers his quantity (taxes) high, while the customer prefers his (revenues) low.

[11] This maximization of cash flows is, for flowthrough, equivalent to maximization of earnings, since the schedule of book depreciation is assumed to be fixed.

and R = revenues. Or, still assuming revenues equal to revenue requirements,

$$C = RR - T.$$

From equations (5) and (6),

$$C_{FT_j} = \rho X_j + D_{B_j}. \tag{24}$$

We notice that C_{FT_j} does not involve D_T; that is, it is independent of the method of tax depreciation and is in fact the same, for each j, for *SL-SL* and for *FT*.

From equations (5b) and (6b)

$$C_{N_j} = \rho Y_j + (1 - T_0)D_{B_j} + T_0 D_{T_j}. \tag{25}$$

Subtracting (25) from (24),

$$\Delta C_j \equiv C_{FT_j} - C_{N_j} = \rho N_j + T_0(D_{B_j} - D_{T_j}).$$

Again using Lemma 1,

$$PW_r(\Delta C) = T_0\left(1 - \frac{\rho}{r}\right)PW_r(D_B - D_T). \tag{26}$$

The solution of $PW_r(\Delta C) = 0$ is

$$r = r_3 = \rho.$$

Equation (26) would at first seem to show that the firm which wishes to maximize the present worth of future cash flows should prefer flowthrough to normalization if its discount rate is less than ρ, be indifferent if it equals ρ, and prefer normalization to flowthrough if it is greater than ρ. Here, however, we have fallen into a logical difficulty: The discount rate of the firm *is* its actual rate of return, say ρ'. Now we cannot have ρ' greater than ρ, the allowed rate of return, because of our assumption of instant and accurate rate-regulation. If ρ' is less than ρ, then the firm is not succeeding in collecting its required revenues—that is, regulation is not binding, and R is not given by Equations (5), (5a), or (5b) for *SL-SL*, *FT*, or *N* respectively. R is now simply the revenues the firm can succeed in collecting, regardless of the treatment of tax depreciation. In this case our assumption of earnings proportional to net plant for each item or vintage is no longer applicable; presumably we would have to shift to the assumption of level revenues over the life of each item, which would take us outside the scope of the present discussion.

We conclude that our model enables us to state only that a firm whose only concern is to maximize the present worth of cash flows, and which is earning its allowed rate of return, should be indifferent as to flowthrough vs. normalization.

☐ **The Investors' Points of View.** A quantity of interest to the equity holder might be return less interest expense, i.e.,

$$(\rho - i\delta)X \quad \text{for } FT,$$
$$(\rho - i\delta)Y \quad \text{for } N.$$

Since $Y_j \leq X_j$ for all j, the equity holder should always prefer flowthrough to normalization, if we assume $RR = R$; that is, if we ignore the higher risk associated with flowthrough, about which we shall have more to say later.

The only contingency which need worry the bondholder in our model is the possibility of default. Hence he will be interested in the ratio of pre-tax earnings to interest expense. The effects of various tax depreciation methods on this quantity are discussed in Section 5.

☐ **The Growing Firm.** So far we have only talked about a single vintage of plant, but our conclusions carry over directly to a company experiencing an arbitrary growth process. This result is expressed by

THEOREM 3: *Theorems 1 and 2 apply also to an arbitrary superposition of vintages, i.e., to a company whose gross additions to plant vary in an arbitrary manner from year to year, given only that the revenues or taxes whose present worths are being calculated apply to a sequence of vintages of plant whose entire lives are included in the interval under consideration.*

Theorem 3 follows from the considerations that

(1) the present-worth operation is additive, i.e., given two sequences (a) and (a'),

$$PW_r[(a) + (a')] = PW_r(a) + PW_r(a'),$$

so that the present worth of a superposition of vintages is the sum of their present worths;

(2) Theorems 1 and 2 apply to each vintage individually if the present-worth operation is referred to the date of installation of that vintage;

(3) for any two (integer) instants t_1, t_2 and any sequence (a), $[PW_r(a)$ referred to $t_1] = (1 + r)^{t_1 - t_2} \cdot [PW_r(a)$ referred to $t_2]$, and multiplication by the geometrical factor cannot change the sign of an inequality.

Since this result is valid only for entire lives, and since life tables may (theoretically) have infinite tails, it is implied that we must consider a time interval extending from the date of installation of the first vintage under consideration forward to infinity.

It should be noted that the sums in question may fail to exist, and in fact do so in the case of uniform, infinitely extended, exponential growth at a rate greater than r. In this case Theorems 1 and 2 have no meaning.[12]

---

[12] The discussion may be clarified by the following formal proof for a specific case.

Suppose that an initial investment of value unity at time zero leads, under two different methods of tax depreciation, to taxes t_j and t'_j in the j'th year of its life, and that

$$PW_r(t) < PW_r(t'). \tag{27}$$

Consider investments of value K_i $(i = 1, 2, \ldots)$ at the beginnings of years i Then the taxes in year k on the plant installed at the start of year i $(k > i)$ are either

$$K_i t_{k-i+1} \text{ or } K_i t'_{k-i+1},$$

and the total taxes in year k are either

$$T_k = \sum_{i=1}^{k} K_i t_{k-i+1} \text{ or } T'_k = \sum_{i=1}^{k} K_i t'_{k-i+1}.$$

We have

$$PW_r(T) = \sum_{k=1}^{\infty} (1 + r)^{-k} \sum_{i=1}^{k} K_i t_{k-i+1} = \sum_{i=1}^{\infty} K_i \sum_{k=i}^{\infty} (1 + r)^{-k} t_{k-i+1}$$

$$= \sum_{i=1}^{\infty} K_i \sum_{k=1}^{\infty} (1 + r)^{k+i-1} t_k = PW_r(t) \sum_{i=1}^{\infty} K_i (1 + r)^{i-1}. \tag{28}$$

Theorem 3 mentions only the signs of the differences of present worths of *RR*, *T*, and *C* for a superposition of vintages. The ratios of these quantities are close to unity, as in the single-vintage cases, for reasonable values of the parameters (see Table 1). In fact these ratios are the same as in the single vintage case, since each quantity for a superposition of vintages is related to the corresponding quantity for a single vintage by a factor which depends only on the growth pattern and not at all on the tax-depreciation method (see equation (29)).

4. Results on rate increases and decreases

■ We now turn our attention to the capital cost per unit of gross plant. We denote gross plant by Z. We feel that the quantity (RR/Z) bears a close relationship to price, since it may be supposed that for a fixed unit price, R is proportional to Z (it is the gross, or physical, plant which is producing the revenues), and for fixed Z, R is proportional to unit price.

The reader will notice several simplifying assumptions in the above paragraph, namely:

(1) we assume constant returns to scale;
(2) we are considering only a single service with a single unit price—alternatively we are dealing only with the average price of a unit of service of any kind, if such a quantity can be defined;
(3) we assume perfectly inelastic demand;
(4) we are still assuming equality between revenue requirements and revenues (*RR* and *R*).

We make these simplifications without much apology; once the implications of the present model are understood, it may be possible to remove the simplifications one by one.

What we are interested in is the variation of (RR/Z) over time as depreciation policies change. We shall perform the analysis for an exponentially growing plant; such growth is not untypical of utilities in recent years. We now shift to a continuous time model for ease of calculation.[13]

□ **Unit Capital Cost as a Function of Time.** Suppose then that the plant has been growing exponentially since the beginning of time and will continue to do so forever. The plant installed in the interval $(w, w + dw)$ is

$$Ce^{\gamma w}dw\,, \quad (-\infty < w < \infty)\,,$$

where C is some positive constant.

Here γ is the continuous growth rate. If g is the annual growth rate, then

$$\gamma = \ln(1 + g)\,.$$

Similarly,

$$PW_r(T') = PW_r(t') \sum_{i=1}^{\infty} K_i(1 + r)^{i-1}. \tag{29}$$

It follows at once from equations (27), (28), and (29) that

$$PW_r(T) < PW_r(T'), \tag{30}$$

which is the desired result.

[13] Thus, from now on all growth rates, discount rates, etc. are to be interpreted in the sense of continuous compounding

Let t be any time. The survivors at t of plant installed in $(w, w + dw)$ for $w < t$ are

$$Ce^{\gamma w}f(t - w)dw\,,$$

$t - w$ being the age. Thus the gross plant at time t is

$$Z(t) = \int_{w=-\infty}^{t} Ce^{\gamma w}dw\ \ (t - w) = Ce^{\gamma t}\phi(\gamma)\,. \qquad (31)$$

Here we have introduced the Laplace transform,

$$\phi(\gamma) = \int_{x=0}^{\infty} e^{-\gamma x}f(x)dx\,.$$

Suppose that until $t = 0$ straight-line depreciation has been used for both book and tax purposes, but that at $t = 0$ accelerated depreciation is adopted for tax purposes, on new plant only. Suppose further that the vintage-group method is used for both book and tax.

Let us introduce the new notation $d_B(y) =$ book depreciation rate at age y of plant of initial value one unit, and let $d_T(y)$ be similarly defined for tax depreciation.

The total book depreciation at time t is

$$D_B(t) = \int_{w=-\infty}^{t} Ce^{\gamma w}dw\cdot d_B(t - w) = Ce^{\gamma t}\delta_B(\gamma)\,, \qquad (32)$$

where $\delta_B(\gamma)$ is the Laplace transform of $d_B(y)$.

The total tax depreciation at time t is

$$D_T(t) = \int_{w=-\infty}^{0} Ce^{\gamma w}dw\cdot d_B(t - w) + \int_{w=0}^{t} Ce^{\gamma w}dw\cdot d_T(t - w)\,,$$

where the second integral is defined to be zero for $t < 0$. In other words,

$$D_T(t) = \int_{w=-\infty}^{t} Ce^{\gamma w}dw\cdot d_B(t - w) + \int_{w=0}^{t} Ce^{\gamma w}dw[d_T(t - w) - d_B(t - w)]\,,$$

or

$$D_T(t) = Ce^{\gamma t}\delta_B(\gamma) + Ce^{\gamma t}\int_{w=0}^{t} e^{-\gamma w}[d_T(w) - d_B(w)]dw\,. \qquad (33)$$

Let us now calculate the total net plant, $X(t)$, at time t. For plant of initial value one unit, the net plant at age y is

$$1 - \int_{w=0}^{y} d_B(u)du\,.$$

Therefore

$$X(t) = \int_{w=-\infty}^{t} Ce^{\gamma w}dw\left[1 - \int_{u=0}^{t-w} d_B(u)du\right].$$

After some transformation, we find

$$X(t) = \frac{Ce^{\gamma t}}{\gamma}[1 - \delta_B(\gamma)]. \tag{34}$$

We are now in a position to express $RR(t)$ before and after a switch to accelerated tax depreciation with flowthrough. Substituting equations (32), (33), (34) in equation (5), we find

$$\begin{aligned} RR(t) = \frac{1}{1 - T_0} Ce^{\gamma t} \Bigg\{ & \frac{\rho - T_0 i\delta}{\gamma}[1 - \delta_B(\gamma)] + (1 - T_0)\delta_B(\gamma) \\ & - T_0 \int_{w=0}^{t} e^{-\gamma w}[d_T(w) - d_B(w)]dw \Bigg\}. \end{aligned} \tag{35}$$

Now dividing by equation (31),

$$\begin{aligned} \frac{RR(t)}{Z(t)} = \frac{1}{1 - T_0} \cdot \frac{1}{\phi(\gamma)} \Bigg\{ & \frac{\rho - T_0 i\delta}{\gamma}[1 - \delta_B(\gamma)] + (1 - T_0)\delta_B(\gamma) \\ & - T_0 \int_{w=0}^{t} e^{-\gamma w}[d_T(w) - d_B(w)]dw \Bigg\}. \end{aligned} \tag{36}$$

One simplification can be made; since, for a utility, book depreciation must be straight-line, we have

$$d_B(t) = \frac{1}{L} f(t),$$

so that

$$\delta_B(\gamma) = \frac{1}{L}\phi(\gamma). \tag{37}$$

Thus the unit costs are,

$$\begin{aligned} \frac{RR(t)}{Z(t)} = \frac{1}{L} + \frac{1}{1 - T_0} \cdot \frac{1}{\phi(\gamma)} \Bigg\{ & \frac{\rho - T_0 i\delta}{\gamma}\left[1 - \frac{\phi(\gamma)}{L}\right] \\ & - T_0 \int_{w=0}^{t} e^{-\gamma w}[d_T(w) - d_B(w)]dw \Bigg\}. \end{aligned} \tag{38}$$

Several points about equation (38) are worth noting:

1) Only one term depends on t, namely the subtractive integral. This term is nonexistent before the switch to flowthrough at time $t = 0$. Thus:

THEOREM 4: *Unit capital costs when SL depreciation is used for book and tax are constant and are given by*

$$\frac{RR}{Z} = \frac{1}{L} + \frac{1}{1 - T_0} \cdot \frac{1}{\phi(\gamma)} \cdot \frac{\rho - T_0 i\delta}{\gamma}\left[1 - \frac{\phi(\gamma)}{L}\right]. \tag{39}$$

2) The subtractive term in equation (38) at first increases after the switch, in the years during which $d_T(t)$ for the first vintage of new

flowthrough plant is greater than $d_B(t)$, and then decreases after the depreciation schedules cross. This turnaround occurs at an instant, t_1, which is independent of the growth rate and approximately equal to $L/2$.[14] Following the turnaround, capital costs continue to increase, approaching, as $t \to \infty$, the limiting value:

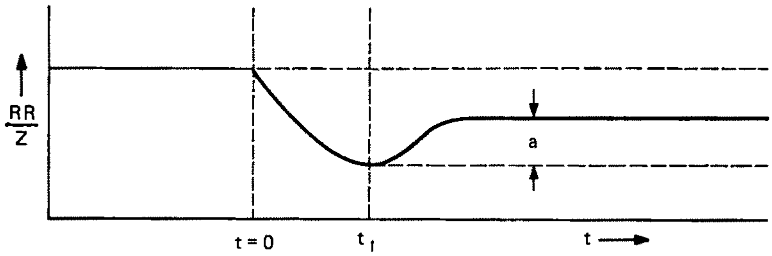
$$\frac{RR}{Z} = \frac{1}{L} + \frac{1}{1 - T_0} \cdot \frac{1}{\phi(\gamma)} \times \left\{ \frac{\rho - T_0 i \mathcal{E}}{\gamma} \left[1 - \frac{\phi(\gamma)}{L} \right] - T_0[\delta_T(\gamma) - \delta_B(\gamma)] \right\}. \quad (40)$$

It is apparent that for zero growth rate, we have $\delta_T(0) = \delta_B(0) = 1$, so that RR/Z ultimately rises back to the *SL-SL* value.[15] For any positive growth rate, the limiting value is less than this. For extremely large growth rates, the contributions to the indefinite integral after the turnaround are negligible, so that the increase in RR/Z after the turnaround is negligible.

All this may be summarized in

THEOREM 5: *The pattern of unit capital costs, before and after a switch to accelerated tax depreciation with flowthrough, has the general shape shown in Figure 2.*

FIGURE 2



Here the interval t_1 is independent of the growth rate and is of the order of one-half the mean life of plant, while the increase, a, generally decreases with increasing growth rate; for $\gamma = 0$, a is such as to bring RR/Z back up to its *SL-SL* value; as $\gamma \to \infty$, $a \to 0$.

Regulatory Interpretation. Theorem 5 shows us the sense in which the assertion—that the risk of flowthrough for a growing company is without force—is true, and the sense in which it is not. It is true that unit capital costs, and hence prices, *ceteris paribus*, after a switch to flowthrough, are always less than they would be under *SL-SL*. On the other hand, we may assume that during the period $(0, t_1)$ rate reductions will be imposed by the commission; if they are not, the rate of return will rise above its allowed value, ρ. But in the case of a switch from *SL-SL* to flowthrough, both the company and the commission should be aware that rate increases will be required, after a fixed interval independent of the growth-rate. (Moreover, to maintain the rate of return the elasticity of the market must be such that these rate increases produce the intended revenue increases.)

---

[14] For *DDB* or *SYD* tax depreciation, and for plausible life tables, t_1 is of the order of $L/2$. For example, for *DDB*, t_1 is given by $0.35L$ for a square life table and by $0.69L$ for an exponential life table These may be taken for practical purposes as limiting cases. For the #2 life table mentioned in the numerical example below, t_1 is $0.43L$. For bizarre life tables (e.g., those in which 90% of the plant dies in the first 10% of the mean life and the remainder is extremely long-lived), t_1/L may be larger. For *ELG* (as opposed to the *VG* method being discussed here), t_1/L seems to tend to be smaller. The main point remains that t_1/L is a constant, independent of the growth rate, and is of the order of $\frac{1}{2}$ for all plausible cases.

[15] Although γ appears in the denominator, the limit as $\gamma \to 0$ of (RR/Z) is finite.

☐ **Total Rate Increases Required.** It should be observed that although the *price* increase, a, is small for large γ, the total rate (revenue requirements) increase in dollars need not be; this is because for large γ, Z is increasing rapidly, and thus a small increase in (RR/Z) may correspond to a large increase in RR. More precisely, suppose the revenue requirements at time t_2 and $t_3 (t_1 < t_2 < t_3)$ have the values RR_2 and RR_3. Without price changes the revenues at t_3 will be, according to our hypotheses, $RR_2\,(Z_3/Z_2)$. Thus the rate increase (dollar revenue increase above that provided by growth) required in this interval is

$$RR_3 - RR_2\left(\frac{Z_3}{Z_2}\right) = Z_3\left[\left(\frac{RR}{Z}\right)_3 - \left(\frac{RR}{Z}\right)_2\right]. \qquad (41)$$

The continuous analogue of equation (41) states that the rate increase required in dt is

$$Z(t)\frac{d}{dt}\left(\frac{RR}{Z}\right)dt. \qquad (42)$$

Therefore, the total rate increase required after the turnaround time t_1 is

$$\int_{t=t_1}^{\infty} Z(t)d(RR/Z). \qquad (43)$$

Using equations (31) and (38) we see that this is

$$-\int_{t=t_1}^{\infty} C\phi(\gamma)e^{\gamma t}T_0e^{-\gamma t}[d_T(t) - d_B(t)]dt.$$

Thus we obtain

THEOREM 6: *For exponentially growing plant of value $C\phi(\gamma)$ at $t = 0$, the total dollar rate increases required under flowthrough after the turnaround point, t_1, are*

$$C\phi(\gamma)T_0\int_{t=t_1}^{\infty} [d_B(t) - d_T(t)]dt. \qquad (44)$$

Note that, given the gross plant at $t = 0$, this quantity is independent of the growth rate. It is interesting that for life tables which depend only on (t/L), which is commonly the case, the integral is also independent of L.

Figures 3 and 4 are computer-generated graphs of RR/Z for growth rates of 5 percent and 12 percent annually. The other

FIGURE 3

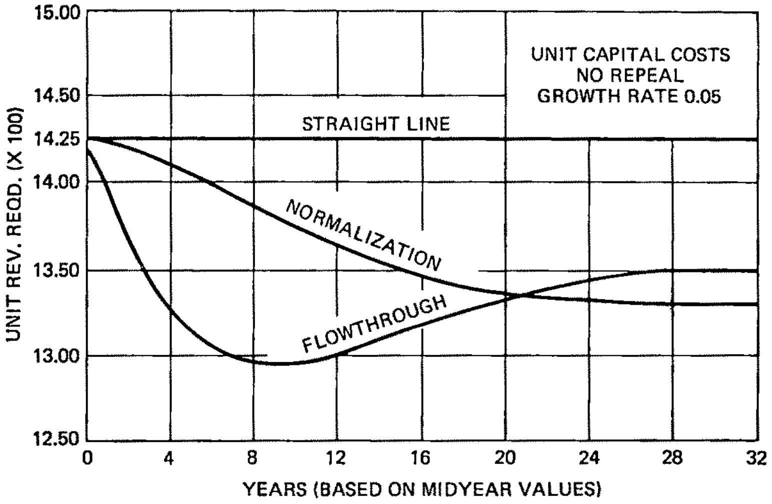
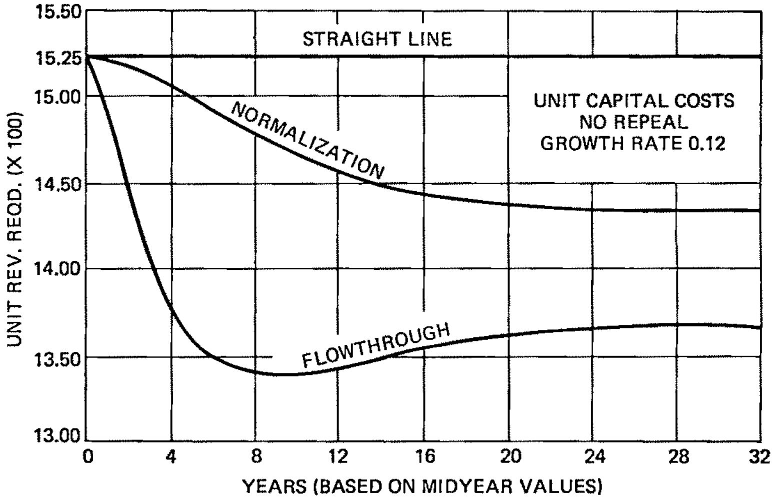


FIGURE 4



parameters assumed are

$$T_0 = 0.5$$
$$\rho = 0.08$$
$$i = 0.06$$
$$\delta = 0.5$$
$$L = 20$$

We assume *SL* for book, *DDB* for tax. It is also assumed that the life table has the form

$$f(t) = (1.02808)^t \times 0.46334^{(1.0499994^t - 1)},$$

a shape which is characteristic of some types of utility plant. (This is called a #2 life table, and is one of a family which, by variation of the parameters, can range from an almost square to an almost exponential shape.)

In these examples the turnaround occurs at $t_1 = 10$ years after the switch to accelerated depreciation with flowthrough, and, if we further assume a gross investment at $t = 0$ of \$1,000,000, the rate increases required starting at t_1 total, in both cases, \$87,300. It can be seen that for a real utility, e.g., AT&T, these are substantial indeed (about 4 billion dollars, if AT&T were to switch to *FT* in 1970).

Figures 3 and 4 also include curves of (RR/Z) for normalization.[16]

We note the following consequences of equation (48):

1) The only time-dependent term is the subtractive integral. This has been shown to be monotone increasing.
Thus

THEOREM 7: *After a switch to accelerated tax depreciation with normalization, rate cuts will be imposed, but no turnaround occurs; rate increases are never required.*[17]

2) The limiting value given by equation (48) as $t \to \infty$ is

$$\frac{RR}{Z} = \frac{1}{L} + \frac{1}{1 - T_0}\,\frac{\rho - T_0 i\delta}{\gamma\phi(\gamma)}\left\{1 - \frac{\phi(\gamma)}{L} - T_0[\delta_T(\gamma) - \delta_B(\gamma)]\right\}. \quad (49)$$

Comparison of equations (40) and (49) gives us

THEOREM 8: *There is a critical growth-rate*

$$\gamma_1 = \rho - T_0 i\delta\,, \quad (50)$$

such that for $\gamma > \gamma_1$ the limiting value of unit capital costs (as $t \to \infty$) is greater for normalization than for flowthrough, while for $\gamma < \gamma_1$ the contrary is the case.

It will be observed that the critical growth-rate, γ_1, is the same as the critical discount-rate mentioned in Theorem 1. There may at first seem to be a contradiction: if $\gamma > \gamma_1$, then (RR/Z) is at all times greater for normalization than for flowthrough, and so is RR;

---

[16] In order to perform the comparable calculations for normalization, we must use equation (5b), and hence need an expression for $Y(t)$ for exponentially growing plant. We know that $Y = X - N$, and we have equation (34) for X, so we need a value for $N(t)$ for exponentially growing plant. For plant installed before the switch, N is of course zero. For plant installed in the interval $(w, w + dw)$, where $w > 0$, the normalization reserve at time $t > w$ is

$$Ce^{\gamma w}dw\,T_0\int_{u=0}^{t-w}[d_T(u) - d_B(u)]\,du$$

Thus the total normalization reserve at time t is,

$$N(t) = T_0\int_{w=0}^{t} Ce^{\gamma w}dw\int_{u=0}^{t-w}[d_T(u) - d_B(u)]\,du\,.$$

After some manipulation, we find

$$N(t) = \frac{T_0 Ce^{\gamma t}}{\gamma}\int_{w=0}^{t}(e^{-\gamma w} - e^{-\gamma t})[d_T(w) - d_B(w)]\,dw\,. \quad (45)$$

Note that

$$\frac{N(t)}{Z(t)} = \frac{T_0}{\gamma\phi(\gamma)}\int_{w=0}^{t}(e^{-\gamma w} - e^{-\gamma t})[d_T(w) - d_B(w)]\,dw\,. \quad (46)$$

For functions $d_B(\ )$ and $d_T(\ )$ which have only one crossing (at the instant we have been calling t_1), this quantity is monotone increasing with time

$$\frac{d}{dt}\left[\frac{N(t)}{Z(t)}\right] = \frac{T_0 e^{-\gamma t}}{\phi(\gamma)}\int_{w=0}^{t}[d_T(w) - d_B(w)]\,dw > 0 \quad \text{for} \quad t > 0\,. \quad (47)$$

The expression for unit capital costs in the case of normalization is

$$\frac{RR}{Z} = \frac{1}{L} + \frac{1}{1 - T_0}\,\frac{\rho - T_0 i\delta}{\gamma\phi(\gamma)}\left\{1 - \frac{\phi(\gamma)}{L} - T_0\int_{w=0}^{t}(e^{-\gamma w} - e^{-\gamma t})[d_T(w) - d_B(w)]\,dw\right\}. \quad (48)$$

[17] For the #2 life table referred to above, and for *DDB* tax depreciation, there is a second crossing of $D_B(\ )$ and $d_T(\ )$ far out on the tail of the life table. This leads to a slight violation of Theorem 7, i.e., to a slight increase in (RR/Z), unless we shift back from *DDB* to *SL*-remaining-life for tax purposes, as described in Appendix 1.

FIGURE 5

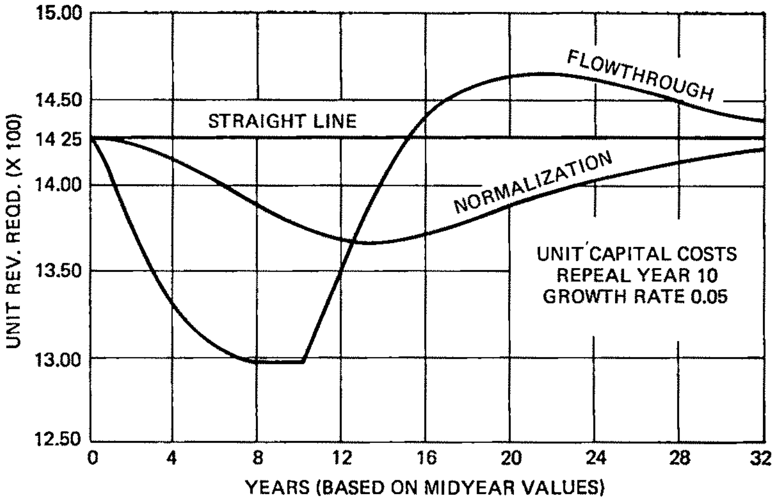
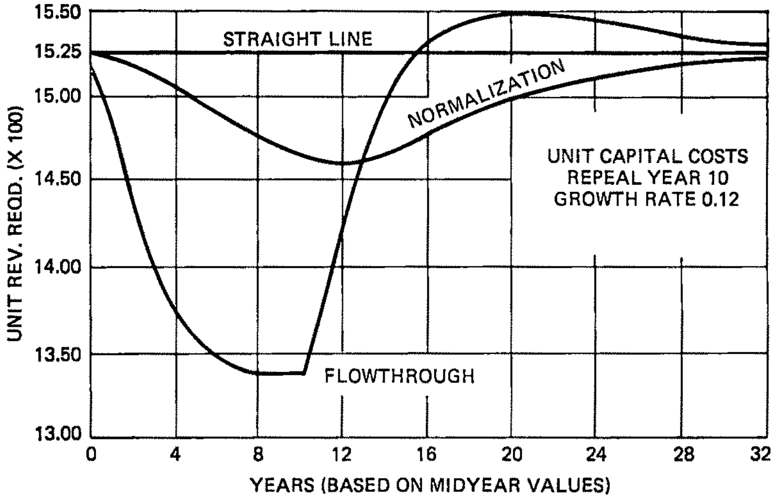


FIGURE 6



how then can PW_r (RR_N) be less than PW_r (RR_{FT}), even if $r < r_1$? The answer is that for $r < r_1 = \gamma_1 < \gamma$, the present worths in question do not exist, and hence Theorem 1 is without meaning.

☐ **Repeal.** By methods similar to the above, one can derive expressions for (RR/Z) in case the firm switches to accelerated tax depreciation at $t = 0$ and then back to *SL-SL*, on new plant only, at some later time, as would be the case if permission to use accelerated depreciation were to be repealed.

Figures 5 and 6 are graphs of (RR/Z) in the case of repeal 10 years after the switch, for the two examples used above. We note that:

1) The *FT* curves rise above the *SL-SL* level. Deferred taxes on *FT* vintages must be collected from the customers, and there is no new low-cost *FT* plant to offset them.

For the examples illustrated in Figures 5 and 6, the total rate increases after repeal are of the order of \$28,500 and \$86,200, respectively, per million dollars of plant at $t = 0$. These increases must be effected within 10 years.

2) Even the normalization curves show a moderate upward turn after repeal, although they never exceed the *SL-SL* level. This is because the savings, attributable to the fact that the normalization reserve is not included in the rate base, gradually disappear as normalization plant is replaced by *SL-SL* plant.

We have now come to the end of the principal train of thought of this paper, which seeks to describe the influence of various methods of tax depreciation on the evolution in time of revenue requirements. The formalism we have developed, however, can now provide, with very little additional effort, results concerning two subjects of some practical interest, namely, the effect of tax depreciation on bond ratings and on tax-free dividends.

Some simple considerations on these subjects are given in the following section.

5. Results on times interest earned and tax-free dividends

□ **Interest Coverage.** One of the important quantities used in rating bonds is the interest coverage or "times interest earned"; this is the ratio of pre-tax earnings to interest expense, and is taken as some sort of inverse measure of the risk that earnings will fail to cover interest expense, even though this risk is usually very small. Pre-tax earnings are used because taxes are computed after subtraction of interest expense. We shall denote this ratio by A.

For *SL-SL*,

$$A = \frac{\rho X + T}{i\delta X}, \tag{51}$$

or, using equation (6a),

$$A = \frac{1}{1 - T_0}\cdot\frac{\rho - T_0 i\delta}{i\delta}. \tag{52}$$

For *FT*, we substitute equation (6) in equation (51) and obtain

$$A = \frac{1}{1 - T_0}\cdot\frac{\rho - T_0 i\delta}{i\delta} + \frac{T_0}{1 - T_0}\cdot\frac{D_B - D_T}{i\delta X}. \tag{53}$$

Since with positive growth D_B on total plant is less than D_T, *FT* reduces interest coverage.

In the case of normalization, the revenues collected to add to the normalization reserve are included in the pre-tax earnings. The interest expense is computed on the reduced rate-base Y. Thus for normalization

$$A = \frac{\rho Y + T + T_0(D_T - D_B)}{i\delta Y}. \tag{54}$$

Using equation (6b),

$$A = \frac{1}{1 - T_0}\,\frac{\rho - T_0 i\delta}{i\delta}, \tag{55}$$

which is identical with equation (52). All this may be summarized in

THEOREM 9: *A switch to accelerated tax depreciation with flowthrough reduces interest coverage. A switch to accelerated tax depreciation with normalization has no effect on interest coverage.*

Tables 2 and 3 show the evolution of interest coverage after a switch to accelerated tax depreciation for the two examples used previously.

TABLE 2

INTEREST COVERAGE

NO REPEAL

GROWTH RATE 0.05

YEAR	SL-SL	SL-DDB-FT	SL-DDB-N
0	4.33	4.24	4 33
1	4.33	4.08	4 33
2	4.33	3.97	4 33
3	4.33	3 88	4 33
4	4.33	3.81	4.33
5	4.33	3.77	4.33
6	4.33	3.74	4.33
7	4.33	3.72	4 33
8	4.33	3.72	4 33
9	4.33	3.72	4.33
10	4.33	3 73	4.33
11	4.33	3 75	4.33
12	4.33	3.76	4.33
13	4.33	3.78	4.33
14	4.33	3 80	4.33
15	4.33	3.82	4.33
16	4.33	3 84	4 33
17	4.33	3.86	4.33
18	4.33	3.88	4.33
19	4.33	3.89	4.33
20	4.33	3.91	4.33
21	4.33	3.92	4.33
22	4.33	3.93	4.33
23	4.33	3.94	4.33
24	4.33	3.95	4.33
25	4.33	3.95	4.33
26	4.33	3.96	4.33
27	4.33	3.97	4.33
28	4.33	3.97	4.33
29	4.33	3.97	4.33
30	4.33	3 98	4.33
31	4.33	3.98	4.33

In the case, not included in this model, that a firm uses not only an accelerated depreciation schedule but also a shortened mean life for tax purposes (generally through the use of "guideline" tax depreciation), the effect of flowthrough on interest coverage can be quite drastic.

☐ **Tax-Free Dividends.** Tax-free dividends arise in the following way: if the firm uses different depreciation methods for book and tax purposes, it may be thought of as keeping two sets of books. Since one of its expenses, namely depreciation, is higher on its tax books, its net earnings after interest ("tax earnings and profits") are lower on its tax books. It may happen that these current tax earnings and profits, plus tax earnings and profits accumulated over prior years, are insufficient to cover dividends. To the extent that this happens, the dividend payments are judged to be a return of capital to the investor, and he does not pay taxes on them.

Tax earnings and profits are given by

$$B = RR - T - D_T - i\delta X \tag{56}$$

(again assuming $RR = R$). From equation (2),

$$B = (\rho - i\delta)X + D_B - D_T .$$

TABLE 3

INTEREST COVERAGE

NO REPEAL

GROWTH RATE 0.12

YEAR	SL-SL	SL-DDB-FT	SL-DDB-N
0	4.33	4.20	4.33
1	4.33	3.98	4.33
2	4.33	3.82	4.33
3	4.33	3.71	4.33
4	4.33	3.64	4.33
5	4.33	3.59	4.33
6	4.33	3.56	4.33
7	4.33	3.55	4.33
8	4.33	3.55	4.33
9	4.33	3.55	4.33
10	4.33	3.56	4.33
11	4.33	3.57	4.33
12	4.33	3.58	4.33
13	4.33	3.59	4.33
14	4.33	3.60	4.33
15	4.33	3.61	4.33
16	4.33	3.62	4.33
17	4.33	3.63	4.33
18	4.33	3.64	4.33
19	4.33	3.65	4.33
20	4.33	3.65	4.33
21	4.33	3.66	4.33
22	4.33	3.66	4.33
23	4.33	3.66	4.33
24	4.33	3.67	4.33
25	4.33	3.67	4.33
26	4.33	3.67	4.33
27	4.33	3.67	4.33
28	4.33	3.67	4.33
29	4.33	3.67	4.33
30	4.33	3.67	4.33
31	4.33	3.68	4.33

Let the payout ratio be p. Then the dividend expense is

$$p(\rho - i\delta)X .$$

Assuming for simplicity that tax surplus is already zero, tax-free dividends will arise if

$$p(\rho - i\delta)X > (\rho - i\delta)X + D_B - D_T ,$$

and in this case their dollar amount is

$$p(\rho - i\delta)X - [(\rho - i\delta)X + D_B - D_T] ,$$

or

$$\text{tax-free dividends} = -(1 - p)(\rho - i\delta)X + D_T - D_B . \tag{57}$$

(We have not included consideration of dividends on preferred stock; with some utilities these are at present partially tax free; dividends on common stock are totally so.)

The situation with normalization is somewhat different. Here

$$B = (\rho - i\delta)Y + (1 - T_0)(D_B - D_T), \tag{58}$$

and

$$\text{tax-free dividends} = -(1 - p)(\rho - i\delta)Y + (1 - T_0)(D_T - D_B). \tag{59}$$

For exponential growth, long after the switch to accelerated tax depreciation, we obtain by substitution of equations (32), (33), and (45), that the excess of tax-free dividends under flowthrough over those under normalization is:

$$T_0[\delta_T(\gamma) - \delta_B(\gamma)]\cdot\left[1 - \frac{(1 - p)(\rho - i\delta)}{\gamma}\right]\cdot Ce^{\gamma t}. \tag{60}$$

Now for any positive growth rate, $\delta_T(\gamma) > \delta_B(\gamma)$. The second square bracket is also likely to be positive. For example, if

$$p = 0.6,$$
$$\gamma = 0.08,$$
$$\delta = 0.5,$$
$$i = 0.05,$$
$$T_0 = 0.5,$$

then the last factor is 0.625. Thus it can be seen that for normalization tax-free dividends are likely to be smaller and set in later than for flowthrough.[18]

6. Summary of results

■ We now summarize our conclusions. It will be understood that propositions stated here were derived within the framework of the model and with conditions and qualifications detailed in the text.

1. For a single vintage of plant, the rational customer (one who prefers smaller present worth of payments) should favor normalization over flowthrough if his personal discount rate is less than $r_1 = \rho - T_0 i\delta$ (the "tax-adjusted rate of return"), be indifferent if it equals r_1, and favor flowthrough if it exceeds r_1. The numerical differences involved are small for reasonable discount rates.
2. For a single vintage of plant, the rational tax collector should favor flowthrough over normalization if the relevant discount rate is less than $r_2 = \rho - i\delta$ (the unit cost of equity), be indifferent if it equals r_2, and favor normalization if it exceeds r_2.
3. Results 1 and 2 apply also to an arbitrary superposition of vintages, i.e., to a growing firm.
4. A regulated utility which switches to accelerated tax depreciation with flowthrough will at first experience decreasing unit capital costs, and hence rate cuts; after an interval, however, capital costs will increase, and hence rate increases will be required to maintain the rate of return. This interval is independent of the growth-rate

---

[18] It is interesting to note that under the new tax law the possibility of tax-free dividends will gradually disappear, only SL depreciation will be allowed for computing tax earnings and profits.

and is of the order of one-half the mean life of plant. The greater the growth rate, however, the smaller the increase in unit capital costs.

5. The total dollar rate increases required in this situation are independent of growth rate and mean life.

6. A regulated utility that switches to accelerated tax depreciation with normalization will experience decreasing unit capital costs, and hence rate cuts, but will not require rate increases.

7. For growth rates less than the critical value $\gamma_1 = \rho - T_0 i\delta$ (numerically equal to the "tax-adjusted rate of return"), unit capital costs will ultimately stabilize at a value greater for flowthrough than for normalization; for growth-rates greater than γ_1, the contrary is the case.

8. A switch to accelerated tax depreciation with flowthrough reduces "times interest earned." A switch to accelerated tax depreciation with normalization has no effect on this quantity.

9. Flowthrough can, under certain circumstances, lead to tax-free dividends. Normalization is less likely to do so.

Appendix 1

Accelerated depreciation applied to dispersed life table

■ We give in this appendix a precise description of how the *VG* and *ELG* methods are applied to a dispersed life table, for *SL*, *DDB*, and *SYD* depreciation, since a discussion of this rather confusing subject seems not to be generally available. *ELG*, although less commonly used, is more rational; we will discuss it first.

Consider a vintage of plant with life table $f(t)$, mean life L, and representing an initial investment K. The amount of plant of life between x and $x + dx$ is

$$K[f(x) - f(x + dx)]dx.$$

This segment of plant may be regarded as a small vintage with a square life table and life approximately x. Hence, under the *ELG*, method, its *SL* depreciation rate at time t is

$$\begin{cases} -Kf'(x)\dfrac{dx}{x}, & 0 \le t < x, \\ 0, & t > x. \end{cases}$$

The total *ELG-SL* depreciation rate at time t is

$$D(t) = -K\int_{x=t}^{\infty} f'(x)\frac{dx}{x}. \qquad \text{(A1)}$$

The *ELG* method has simply reduced the problem of a dispersed life table to the previous problem of a square life table.

Under *ELG-DDB*, the depreciation rate at time t of a segment of plant of life between x and $(x + dx)$ is

$$-2Kf'(x)e^{-2t/x}\frac{dx}{x}, \quad 0 \le t,$$

so that the total *ELG-DDB* depreciation rate at time t is

$$D(t) = -2K\int_{x=0}^{\infty} \frac{f'(x)}{x} e^{-2t/x}dx. \qquad \text{(A2)}$$

Under *ELG-SYD*, the depreciation rate at time t of a segment of plant of life approximately x is

$$\begin{cases} -Kf'(x)\dfrac{2(x-t)}{x}\cdot\dfrac{dx}{x}, & 0 \leq t \leq x \\ 0, & t > x, \end{cases}$$

so that the total *ELG-DDB* depreciation rate at time t is

$$D(t) = -K\int_{x=t}^{\infty} f'(x)\frac{2(x-t)dx}{x^2}. \tag{A3}$$

The situation in which the *VG* method of applying the depreciation method to the life table is used is computationally simpler but conceptually more obscure.

Under *VG*, each segment of the vintage is treated as if it had the mean life L. Thus under *VG-SL*, the total depreciation rate at time t is

$$D(t) = \frac{K}{L}f(t). \tag{A4}$$

Under *VG-DDB*, the depreciation rate $2/L$ is always applied to the total net plant, so that the total depreciation rate at time t is

$$D(t) = \frac{2K}{L}e^{-2t/L}. \tag{A5}$$

In this case, the depreciation rate depends only on the mean life and not at all on the shape of the life table.

The *VG-SYD* situation is somewhat strange. If we were to stick to equation (16a), we would depreciate all the plant by age L, when much of it may still be in service. To get around this difficulty, we introduce the idea of the mean remaining life at age t. This is the conditional expectation

$$L(t) \equiv E\{\chi - t \mid \chi > t\},$$

where χ is again the random variable representing the life. $L(t)$ is given by

$$L(t) = \int_{w=t}^{\infty} f(x)dx/f(t). \tag{A6}$$

Of course $L(0) = L$.

At time t, the *VG-SYD* depreciation rate is given by the initial *SYD* rate for plant of life $L(t)$, applied to the net plant, i.e., by equation (16a) with L replaced by $L(t)$, t replaced by 0, and K replaced by $f(t)$. This yields, after some manipulation,

$$D(t) = \frac{2K}{L(t)}\exp\left(-2\int_{x=0}^{t}\frac{dx}{L(x)}\right). \tag{A7}$$

It can be verified that in all the above cases the total depreciation over the life of the plant adds up to the initial investment

$$\int_{t=0}^{\infty} D(t)dt = K.$$

There is one further complication in the case of *DDB*. Here it is permitted to switch back to *SL*, based on the remaining life, whenever this procedure would give more rapid depreciation than *DDB*. For example, for the case of *VG-DDB* (equation A5), if one were to switch to *VG-SL* remaining life at time t, the depreciation rate from then on would be

$$D(x) = \frac{X(t)}{f(t)} \cdot \frac{f(x)}{L(t)}, \quad x > t, \tag{A8}$$

that is depreciation would again follow the shape of the life table and be applied to the undepreciated plant at time t. The switch from (A5) to (A8) would be made at the first instant t such that

$$\frac{X(t)}{f(t)} \cdot \frac{f(t)}{L(t)} > \frac{2K}{L} e^{-2t/L}.$$

Since

$$X(t) = Ke^{-2t/L},$$

this simplifies: the instant, t, of the switch is the solution of

$$L(t) = \frac{L}{2}. \tag{A9}$$

Appendix 2

Glossary of principal symbols and abbreviations

■ *Symbols*

A = interest-coverage
B = tax earnings and profits
C = cash-flow = revenue less taxes
δ = debt-ratio
D_B = book depreciation
d_B = book depreciation per unit of initial investment
δ_B = Laplace transform of d_B
D_T = tax depreciation
d_T = tax depreciation per unit of initial investment
δ_T = Laplace transform of d_T
$f(t)$ = life table = fraction of survivors of any given vintage at age t
ϕ = Laplace transform of f
g = annual growth-rate
$g(t)$ = $1-f(t)$ = cumulative distribution of life of plant = $P_r\{X \leq t\}$
γ = continuous growth-rate = $\ln(1+g)$
i = imbedded cost of debt
K = initial investment in a vintage of plant
L = mean life of plant
N = normalization-reserve
P = payout-ratio
PW = present worth
$PW_r(a)$ = present worth, at start of year 1, of a sequence $a_1, a_2, \ldots,$ at ends of years 1, 2, . . .; discounted at discount rate r.
R = revenues
RR = revenue requirements

r = discount-rate
ρ = rate of return
ρ_e = rate of return on equity
T = tax payments
T_0 = F. I. T. rate
X = net plant = rate-base
χ = life of an item of plant (random variable)
Y = net plant less normalization-reserve = rate-base under normalization
Z = gross plant

Abbreviations

DDB = double declining balance
FT = flowthrough
SL-SL = straight line depreciation for both book and tax purposes
SYD = sum of the years' digits
N = normalization

Note: Both discrete-time and continuous-time notations are used. Typical example:
D_{B_j} = book depreciation in year j (dollars)
$D_B(t)$ = book depreciation rate at time t (dollars/year)

References

1. "Internal Revenue Code, 1954," *United States Code*, Title 26, Section 167.
2. "Statement of Administrative Policy regarding Balance Sheet Treatment of Credit Equivalent to Reduction in Income Taxes," *Securities and Exchange Commission Release No. 85, Accounting Series*, February 29, 1960.
3. American Institute of Certified Public Accounts, *Opinions of the Accounting Principles Board*, No. 11, December 1967.
4. "Uniform System of Accounts for Class A and Class B Telephone Companies," *Federal Communications Commission Rules and Regulations* (Title 47, Telecommunication, Chapter I), Part 31, Section 02-80.
5. "Re The Southern New England Telephone Company," Connecticut Public Utilities Commission Docket No. 10769, April 23, 1969. *Public Utilities Reports (Third Series)*, Vol. 78 (1969), pp. 504-526.
6. "Re the Pacific Telephone and Telegraph Company," California Public Utilities Commission Decision No. 74917, November 6, 1968. *Public Utilities Reports (Third Series)*, Vol. 77 (1969), pp. 1-54.
7. "Treatment of tax depreciation by regulatory agencies," *Hearings before the Committee on Ways and Means, House of Representatives, Ninety-first Congress, First Session, on the subject of Tax Reform*, Part 10 (March 24 and 25, 1969) and Part 11 (March 26 and 27, 1969).

The theory of public utility pricing and its application

R. H. Coase
Professor of Law and Economics
The University of Chicago

In the 1930's and 1940's, the view came to be held that the right policy was to make public utility prices everywhere equal to marginal cost, even where marginal cost was less than average cost and a government subsidy was required to maintain production. This policy proposal had serious weaknesses. It did not take into account the stimulus to correct forecasting of having a subsequent market test whether consumers were willing to pay the total cost; it ignored the probable effects on the administrative structure, with state enterprise superceding private enterprise and centralized operations superceding decentralized operations; it involved a redistribution of income in favor of consumers of products produced in conditions of decreasing costs; it failed to take into account the misallocation of resources resulting from the additional taxation necessitated by the subsidies.

When appraising a policy proposal, it is necessary to discover its total effect. There is a growing realization among economists that this is the correct approach, and it may be that today what advocates of marginal cost pricing have in mind is that the advantages of making price equal to marginal cost should be taken into account—a position with which there would be general agreement. But if prices are not to equal marginal cost, the question of what the structure of rates should be needs to be examined. The treatment of this subject has not gone very far. But this should not obscure the agreement which exists among economists as to how costs should be calculated (which is very different from the way regulatory commissions view costs). The implications of this difference in approach for price determination may be illustrated by examining the situation created by the FCC's "Above 890-megacycle" decision, which exposed the common carriers to expanded competition from private microwave systems, and the response of the AT&T Company introducing the TELPAK® Service.

■ My subject is the theory of public utility pricing that economists have developed, essentially as a branch of price theory, and which is concerned with the principles which should govern the determination of prices in the utility field. It differs from older treatments of the subject, which discussed and tried to make sense out of the pronouncements and policies of the Supreme Court and the regulatory commissions. The modern economist has taken a different approach.

Professor Coase was born in London, England, and studied at the London School of Economics. After graduating in 1931, he held teaching positions at the Dundee School of Economics, the University of Liverpool, and the London School of Economics. He migrated to the United States in 1947, and has held positions at the Universities of Buffalo, Virginia, and Chicago At the University of Chicago, he now holds a joint appointment as Professor of Economics in the Law School and the Graduate School of Business Professor Coase is editor of the *Journal of Law and Economics*

He only refers to the policies of the courts and commissions to condemn them for their economic illiteracy. I shall follow this modern approach both in the kind of theory I discuss and in my attitude toward the commissions.

The present is, I think, a very interesting time to consider this branch of theory. Much of the controversy of the recent past seems to be in process of resolution, but it is becoming increasingly evident that as we surmount one difficulty other problems emerge. We cross one mountain ridge only to find our way barred by others looming just ahead. As a result, if the courts and commissions ever do start asking the right questions, it is by no means clear that we shall have the right answers.

I propose to consider the view, once widespread, that the public utility pricing problem could be solved by adopting the rule that price should be made equal to marginal cost. The controversy which arose out of this suggestion has, I believe, been largely resolved, but it has led us to consider the more general problem of the optimum structure of rates, a problem which has not been much discussed except under the headings of price discrimination and of the problem of charges for "peak load" services.

The discussion in the literature with which I am familiar does not seem to me to have led to an acceptable solution except in the most general terms, but this does not mean that it is not useful so far as it goes. It may be useful, for example, because some of the problems which have not been solved do not arise in the particular case under consideration.

I will end with a comprehensive illustration of these problems, the deliberations of the Federal Communications Commission (the FCC) with respect to the TELPAK® Service of the American Telephone and Telegraph Company. This case is helpful from the point of view of illustrating the theory, partly because many of the complications which might arise in other cases did not arise to any significant extent in this one.

Proposals for marginal cost pricing

■ I should like to start with a personal history. The idea that price should equal marginal cost, even under conditions in which average cost was greater than marginal cost was, I am quite certain, first drawn to my attention by the work of Abba Lerner in the early 1930's at the London School of Economics. This work of Abba Lerner's, with its reference to positive rents and negative rents (the one occurring when marginal cost was greater than average cost and the other when marginal cost was less than average cost), was later to appear in his book, *The Economics of Control.* [8] But parts of it, and particularly those parts which have a bearing on marginal cost pricing, had been worked out much earlier. Lerner was a member of that group which consisted of Robbins, Hayek, Hicks, Allen, Kaldor and others, who, at the London School of Economics, were making great strides at that time in the development of economic theory.

I was, however, associated with another group of economists who, on the commerce side, were working under the leadership of Professor Arnold Plant. The Plant group was keenly interested in cost analysis and pricing problems, the problem of the peak, and so on, and therefore we paid great attention to Lerner's work and to the other theoretical work being done on cost and pricing. But we thought of it in practical terms. We were influenced but not converted by work such as Lerner's. In our discussions we stressed the practical aspects of these theoretical developments. We treated them seriously, but we treated them as ideas to be applied in the real world.

As a result of Plant's influence, we were particularly interested in multi-part pricing. Paine, who was one of Plant's students, wrote in 1937 on this subject, [13] and a later student Arthur Lewis wrote an article in *Economica* on two-part pricing in 1941. [9]

Of course, the notion of marginal cost pricing has deep roots in the literature and the idea was widely dispersed. It is one of the merits of Mrs. Ruggles' excellent survey articles[1] that they bring out the historical sources of the doctrine in the works of Marshall, Pigou, Allyn Young, Pareto, Barone, and Wicksell. But of most of these contributions many of us were unaware. Certainly I was. Mrs. Ruggles starts her survey of modern developments with Hotelling's 1938 article, [7] but though it had an immediate influence in the United states, as Professor Bonbright and others have testified, I am not sure when it came to be read in Britain. It appeared in 1938, but these were troubled times, and from 1939 on most of us were preoccupied with other matters.

Interestingly enough, in the United States a dissent from the view that price should be made equal to marginal cost was expressed in 1941 by Professor Eli Clemens of the University of Maryland. [1] But notwithstanding Clemens, it is clear that professional opinion at that time—the early 1940's—was fast moving toward general acceptance of the principle of marginal cost pricing.

The next major event was the appearance of the Meade-Fleming symposium entitled "Price and Output Policy of State Enterprise."[2] This symposium had appeared as a paper of the economic section of the Cabinet Office as part of a discussion of how state enterprises ought to be run, with no thought of publication. However, Keynes, who was an advisor in the Treasury as well as editor of the *Economic Journal* saw this paper and was so enthusiastic that he published it.

The publication of this paper was, I think, the high point of the doctrine. Marcus Fleming at this time felt able to claim that the proposal for marginal cost pricing "is not I think open to serious criticism."

Immediately after the Meade-Fleming symposium appeared in the *Economic Journal,* I wrote a short note in the same journal expressing dissent. [2] Shortly afterwards I found I had an ally in the economic section of the Cabinet Office in Tom Wilson, who also wrote a note expressing disagreement with the principle of marginal cost pricing. [19]

In the meantime, James Meade had become head of the economic section of the Cabinet Office and Britain had a Labor Government. A paper was prepared by the economic section setting out the policy which it was considered ought to be followed in the nationalized industries, and this included a suggestion for adopting marginal cost pricing. This proposal was not, however, accepted by the Minister concerned, Mr. Herbert Morrison, and marginal cost pricing has played no part in the pricing policies of the nationalized industries. As it happens, pricing policies in the nationalized industries have tended to develop in ways which I find very congenial, and some of the most interesting work of which I know in the field of pricing is being conducted in the nationalized electricity supply industry in Britain. The nationalized industries have in fact followed a completely different line from that suggested by the marginal cost pricing proposal as originally conceived, and in the meantime, of course, enthusiasm in the profession for marginal cost pricing has become less pronounced.

---

[1] See [14] and [15].

[2] See [6] and [10].

The case for marginal cost pricing examined

■ Hotelling's article has been described as the originating article for modern discussion, and I have no doubt that it was in the United States although it was not for many British economists. Nonetheless, it is a very good article with which to start, as it sets out the argument for marginal cost pricing clearly and strongly. A good deal of the argument is mathematical, but in between the mathematics you will find a clear statement of what Hotelling's proposal is and why he supports it. Hotelling describes his purpose as follows:

> In this paper we shall bring down to date in revised form an argument due essentially to the engineer Jules Dupuit, to the effect that the optimum of the general welfare corresponds to the sale of everything at marginal cost.[3]

This opening statement is somewhat misleading, since Dupuit did not in fact propose that price should be made equal to marginal cost. He merely indicated the kind of loss that would be suffered if you did not make price equal to marginal cost, but he also indicated the kind of loss that would be suffered if you did and therefore left the whole question open. This historical slip is of little importance except to suggest that Hotelling was not attuned to the disadvantages of his policy recommendations and therefore failed to see the significance of these remarks of Dupuit.

We should note here the practical meaning which Hotelling gave to his proposal:

> This means that toll bridges which have recently been reintroduced around New York are inefficient reversions, that all taxes on commodities, including sales taxes, are more objectionable than taxes on income, inheritances and the site value of land; and that the latter taxes might well be applied to cover the fixed cost of electric power plants, water works, railroads, and other industries in which the fixed costs are large, so as to reduce to the level of marginal cost the prices charged for the services and products of these industries.
>
> The efficient way to operate a bridge—and the same applies to a railroad or a factory if we neglect the small cost of an additional unit of product or transportation—is to make it free to the public so long at least as the use of it does not increase to a state of overcrowding. A free bridge costs no more to construct than a toll bridge, and costs less to operate; but society, which must pay the cost in some way or other, gets far more benefit from the bridge if it is free, since in this case, it will be more used. Charging a toll, however small, causes some people to waste time and money in going around by longer but cheaper ways, and prevents others from crossing. The higher the toll, the greater is the damage done in this way. . . .[4]

As is evident, the proposal is pretty far-reaching, but of course it has its strength. If additional supplies of a product are made available at a price which is greater than the cost of the additional factors of production needed to produce that additional unit, then some people are likely to spend their time, and money, in ways that bring in a lower return than if these units were made available at marginal cost. If, for example, a group of factors is made available in use *A* at one dollar and in use *B* at fifty cents, people may well choose to express their demands on the market so that the factors are used in use *B*, because by doing this they get the product of the factors plus fifty cents as compared with their employment in use *A*, in which they only get the product of the factors. This is true, notwithstanding that they value the product of the factors more in use *A* than in use *B*.

---

[3] See [7].

[4] See [7].

Since there are normally in our economic system uses in which factors of production are made available at marginal cost or close to it, to charge a price in certain uses which is above marginal cost would mean that relatively too many factors would be used in uses in which price is equal to marginal cost and relatively too few in uses in which price is above marginal cost. There are too many in the one use and too few in the other if we have regard to maximizing the value of production. This underproduction in some uses and overproduction in others could be avoided if factors of production were made available in all uses at the same price, but this will only be true if price is equal to marginal cost everywhere.

This is a strong argument and was at any rate sufficiently strong to convince many, if not most, of the best economists that a policy in which price is made equal to marginal cost should be carried into effect. Indeed I was somewhat doubtful whether it was right to call my 1946 article "The Marginal Cost Controversy," [3] since up to that time there really had been little controversy. But I consoled myself by thinking that if there had not been much controversy before, there certainly would be after it appeared, and so of course it proved. There was a whole series of articles attacking me, which appeared in *Economica* and the *Journal of Political Economy* after that date.[5]

Though the argument for marginal cost pricing is strong, it does not represent the whole truth. In fact, it ignores some very important aspects of the question which need to be taken into account in coming to a policy recommendation. It was this belief which led me to write my 1946 article. I wished to show that the argument for marginal cost pricing was invalid, but because I held a minority position and it would be assumed that I was wrong, I wished to develop an argument which would show that the case for marginal cost pricing was not inevitably correct. The proposal I was attacking was that prices should be made equal to marginal cost in industries in which average cost was decreasing with increases in output, and in which, therefore, marginal cost was less than average cost. Of the costs incurred in such an industry, it was proposed that an amount equal to marginal cost multiplied by output would be paid for by consumers, but the difference between marginal cost and average cost multiplied by output would be paid for by the Government out of the proceeds of taxation.

I need not bother with the details of the example I used to illustrate my argument. But, given my background, it was not surprising that I started off by considering the possibility of multi-part pricing.

---

[5] See, e.g., [8], [9], [16].

It is clear that it is desirable for a consumer to be allowed to obtain additional units of production at marginal cost, but equally, it seems to me, it can be argued that a consumer should pay an amount equal to the total cost of supplying him.

A consumer does not only have to decide whether to consume additional units of the product. He also has to decide whether it is worth his while to consume the product at all rather than spend his money in some other direction. This can be discovered if the consumer is asked to pay an amount equal to the total costs of supplying him. I suggested that the one way of doing this would be to have a charge independent of consumption to cover the cost not included in marginal cost. Note that marginal cost pricing makes it impossible for consumers to choose rationally between two alternative uses of the factors that are required for production but do not enter into marginal cost. In the one use, factors of production which do not enter into marginal cost would be free. In another use, providing that they enter into marginal cost in that use, they would have to be paid for.

Apparently what the advocates of marginal cost pricing had in mind was that the Government should estimate for each consumer whether he would be willing to pay a sum of money which would cover the total cost. However, if it is decided that the consumer would have been willing to pay a sum of money equal to the total cost, then—and this strikes me as a very paradoxical feature of this argument—he will not be asked to do so. So the Government would estimate whether a consumer would be willing to pay, and if he is willing to pay, it does not charge him.

I found this a very odd feature. But I do not see how it would be possible for any government, or anyone else for that matter, to make accurate estimates at low cost and without knowledge of what would have happened if consumers had been required to pay the cost. The way we discover whether people are willing to pay something is to ask them to pay it, and if we do not have such a system, it becomes extremely difficult to make estimates of whether they would be willing to pay. As I said at the time, if it were possible to make such estimates, at low cost, with accuracy, and without knowledge of what had happened when people had been asked to pay the costs, then it would seem to me that this would lead not to a modification of the pricing system but to its abolition. There really is no point in having a pricing system if the information which it provides is valueless.

But, of course, such estimates, if made, would in practice be very expensive, and they would be inaccurate, and much waste of resources would result from the kind of procedure envisaged by the advocates of marginal cost pricing.

None of the original advocates of marginal cost pricing seems to have given, in my view, sufficient weight to the stimulus to correct forecasting which comes from having a subsequent market test of whether consumers are willing to pay for the total cost of the product. I do not know how one could discover whether people would be willing to pay for something unless, from time to time, they are put in a position in which they can only obtain it by paying for it. Furthermore, it is easy to see that such estimates will be less carefully made in circumstances in which it was never possible to discover whether the estimate on which the decision was based was correct or not.

This proposal also completely ignores the effect general subsidies are likely to have on the administrative structure. In fact, it is impossible to have a widespread decentralization of operations without having a high degree of financial autonomy for the separate operations. This is one of the reasons, for example, why in Britain the Post Office is now set up as a more or less financially independent organization. Since the Post Office is part of the government machine in any case, the difference may seem merely a matter of form. But this is not so. When the Post Office was not financially autonomous, its receipts went into the Treasury and its expenditures were paid for by the Treasury. Before any expenditures could be made they had to be approved by the Treasury, which inevitably played a part in the administration of the Post Office. Such tendencies are strengthened when there is a subsidy, since the person who makes the payment is almost certain to insist on some degree of control because he wants to keep the subsidy down.

Therefore, I believe that some financial autonomy is a necessary aspect of efficient administration, and this again seems to me an argument which indicates that a general policy of marginal cost pricing would be very bad because it would lead in the end to complete centralization of the administration of public utility industries. This was, I suspect, the basic reason why Herbert Morrison, who was a practical politician and experienced in administration in local government, did not look on the proposal for marginal cost pricing with favour. It is one of the merits of Tom Wilson's *Economic Journal* note of 1945 that he emphasized the effect of adopting marginal cost pricing on the administrative structure, state enterprise superceding private enterprise, and centralized operations superceding decentralized operations. The effect of these changes, as Tom Wilson indicated, would be likely to offset any good that marginal cost pricing might bring.

As I see it, the argument for marginal cost pricing, like many propositions in modern welfare economics, is more concerned with diagrams on a blackboard than with the real effects of such policies on the working of the economic system. I have referred to this type of economics as "blackboard economics" because, although factors are moved around and prices are changed, and some people are taxed and others subsidized, the whole process is one which takes place on the blackboard. This is not the way in which one operates with a social system. All that can be done is to set up a new agency, or change the rules under which an old agency operates, or take some other similar action. All that is possible is to operate on social institutions and to discuss economic policy in a sensible way; it is necessary to consider the effect of changing the social institutions with which we work. Therefore, although modern welfare economics may be a good way of teaching economics, it often does a real disservice when it comes to teaching people what economic policy is really about.

Even if the Government were able to make estimates of individual demands and in fact did so, marginal cost pricing would, in my view, be subject to another objection. The Government is supposed to make an estimate of what consumers would be willing to pay, but these consumers are not in fact asked to pay this sum. This money is then available for these consumers to spend on other products. Consumers who buy products which are produced under conditions of decreasing average cost will, therefore, obtain for a given expenditure of money, products embodying a greater value of factors than those who do not. I can see no reason why this redistribution of income should be insignificant in amount, or, for that matter, in a desirable direction. Hotelling admits this as a possibility but claims that by a subsequent redistribution the situation could be brought about in which everyone was better off than before. He did not explain how this redistribution would be accomplished, but it would obviously be a worse procedure than adopting a multi-part pricing system, which would avoid the necessity for such a subsequent redistribution.

In any case, I cannot understand how ordinary taxation procedures could be used to redistribute income from consumers of products produced under conditions of decreasing average cost to all other consumers. An attempt to do this might be made by means of a tax on the consumption of goods produced under conditions of decreasing average costs, but this would either be equivalent to introducing multi-part pricing if a lump sum tax was levied on consumers, or, if a tax per unit of consumption was imposed, it would bring about that divergence between the amount paid for additional units and marginal cost which it was the object of marginal cost pricing to avoid.

Of course, it is usually assumed that the way the Government will obtain the money necessary to pay the subsidy is not by a tax on the subsidized good but by a general increase in the income tax. If we assume that the form of tax used to make good the loss is to be the income tax, we ought to take into account that income taxes are so framed that marginal units of income are taxed, and therefore an income tax has the same kind of adverse effect as a tax on goods and will produce the same kind of unfortunate consequences as would come about through charging a price which is greater than marginal cost. It will, for example, affect the choice between leisure and work, and probably the choice between investments which are more or less risky, and so on.

All in all, it seems to me that marginal cost pricing brings with it substantial disadvantages, and it is clearly worse than multi-part pricing. But, of course, what advocates of marginal cost pricing had in mind as an alternative was not multi-part pricing but average cost pricing, setting the price equal to average cost and above marginal cost. In this case, the argument for marginal cost pricing is strengthened. The disadvantages of marginal cost pricing, which I have already discussed, remain, but we are now comparing it with another system of pricing which also has disadvantages. I claim that this is the usual state of affairs when we are dealing with economic policies. However, the disadvantages of marginal cost pricing are substantial, and there is no reason to suppose that the disadvantages of average cost pricing will in general be worse. Marginal cost pricing certainly allows a better choice at the margin than average cost pricing, but this disadvantage would be reduced and might be offset if marginal cost pricing involved increased income taxes and it surely would.

It is true that in certain circumstances average cost pricing might result in something which ought to be produced not being produced, because a uniform price would not produce sufficient receipts to cover average cost (although even this argument is more complicated, I think, than it is in the usual formulation). But a government or the organization which was given the task of administering the public utility may make many errors, and this could easily offset any good which such a policy might bring.

Average cost pricing may sometimes prevent something being done which ought to be done, but it is also a means of avoiding errors which would certainly be made if a policy of marginal cost pricing were adopted, and of course there is the redistribution of income which would occur and which could not be rectified without producing the same disadvantages which it is the aim of marginal cost pricing to avoid.

Personally, I believe the disadvantages of marginal cost pricing to be so serious as to make it highly undesirable as a general policy, and here, perhaps for the only time, I find myself on the side of Herbert Morrison. This rejection of marginal cost pricing reflects the view that it is a mistake to concentrate simply on the marginal conditions when examining a policy proposal. It is the total effect (in which what happens at the margin is only one factor) which matters.

The present state of opinion

■ I think that nowadays there is a growing realization among economists that this is the correct approach. But there is one economist, William Vickrey, who was an early advocate of marginal cost pricing and has done excellent work in this area of economics, but who still appears to be a supporter of marginal cost pricing. In an article published in 1948 in *The Journal of Political Economy* [17] Vickrey had emphasized the advantages of marginal cost pricing and had minimized the disadvantages which such a pricing policy might bring.

In a paper published in the 1955 *Papers and Proceedings* of the American Economic Association [18], Vickrey still seems to maintain his hostility to average cost pricing and to the granting of financial autonomy to enterprises running public utilities. But in other respects there is a very considerable change in his viewpoint, or so it appears to me.

Vickrey starts by saying:

> . . . in common with any other theoretical principle, the principle of marginal cost pricing is not in practice to be followed absolutely and at all events but is a principle that is to be followed insofar as this is compatible with other desirable objectives and from which deviations of greater or less magnitude are to be desired when conflicting objectives are considered.[6]

This is of course a statement with which it would be difficult to disagree.

When Vickrey goes on to explain some of these considerations which conflict with the principle of marginal cost pricing, it is difficult for me to restrain a cheer. I will quote him:

> By far the most important of these considerations that conflict with the strict application of marginal cost pricing is the need for revenues. Many of the more extreme advocates of marginal cost pricing for decreasing cost industries seem tacitly to assume that the Government has some perfectly costless and neutral source of revenue that is capable of very substantial expansion without ill effects. Such a state might be approached, for example, if we had an income tax free of its multiple defects, evasion proof, with no marginal cost of administration or

---

[6] See [18], p. 605.

compliance, and including in its base not only money income but all forms of direct income in kind, including an imputed value for leisure. Needless to say this is far from the case.[7]

On the general policy issue, Vickrey states:

> Marginal cost pricing must be regarded not as a mere proposal to lower rates generally below the average cost level but rather as an approach which implies a drastic rearrangement of the patterns and structures of rates. Indeed it is this restructuring of rates that is likely to be the greatest contribution of marginal cost pricing to the improvement of overall efficiency of our economy, while the further gains that might be obtainable from the reduction of rates from a self-sustaining level to a marginal cost level, once the pattern of rates has been made to conform as closely as possible to marginal cost, are likely to be relatively small. The issue is thus not primarily one of subsidized versus non-subsidized operations, though this is still an important issue. The dominant issue is one of whether the pattern of rates should be based on tradition, inertia and happenstance, or whether it is to be developed by a careful weighing of the relevant factors with a view to guiding consumers to make an efficient use of the facilities that are available.[8]

What this seems to indicate is that there is no difference of any substance between the position of someone who, like Vickrey, claims to be a supporter of marginal cost pricing and someone who, like myself, regards himself as an opponent. The only substantial difference seems to be one concerning terminology, and that is hardly worth spending much time on.

I mean by marginal cost pricing a policy in which price is made equal to marginal cost. Vickrey seems to mean by marginal cost pricing a policy in which the advantages of making price equal to marginal cost are taken into account. I once summed up my position in the following words:

> We must recognize, it seems to me, that to make price equal to marginal cost would have undesirable results, and furthermore that not to make prices equal to marginal cost would also have undesirable results.
>
> Insofar as we are concerned with public policy, the question is always one of choosing out of the practical alternatives the one which, on balance, seems to get the best results, which probably means that it is not possible to carry the analysis very far except on an industry-to-industry basis.[9]

I think this leads us, as Vickrey indicated, to a study of the appropriate rate structure, a subject which has traditionally been discussed in connection with price discrimination and the problem of the peak. The only trouble with the treatment of these subjects is that it does not seem to me to have gone very far. But we should not disregard the area of agreement that exists among economists, since it relates to questions of great importance with regard to the pricing of public utility services and embodies a point of view which many regulatory commissions do not accept.

Almost all economists would agree, I think, that it is desirable that prices should in general cover costs. Further, most economists would agree that the costs are a marginal, additional, or incremental concept; that this is the only sense in which you can talk about costs; and that there really cannot be in a strict sense a category of fixed costs. Again, I think most economists would agree that the costs which are relevant for pricing are current or future costs, and that the so-called historical costs can be ignored. A final point on which most agree is that the allocation of joint or common costs between products or services for the purpose of determining prices is without meaning. This does not mean that it is impossible to discuss the question of how the total expenses of an enterprise ought to be defrayed. Nor does it mean that cost conditions play no part in determining the answer to this question. But the determination of the optimum structure of rates is not easy because it involves a number of as yet unresolved questions. Economists have, however, advanced part of the way, and what they have resolved may be useful on occasion, as we shall see, in making decisions on practical policy issues.

---

[7] See [18], p. 607.

[8] See [18], pp. 618–19.

[9] See [5], p. 437.

Costs and prices

■ Let us examine the fundamental question of the nature of cost. The answer which an economist gives to this question, the answer which I consider any sensible person must give to this question, is that the cost of doing anything is what is given to do it. An expansion of the supply of any public utility service is brought about by the employment of resources which would otherwise have been available for the production of other goods and services, and the cost of that additional supply of the public utility service is thus the value of the reduction of output elsewhere which comes about in consequence of the diversion of resources to that industry. I would emphasize that what is lost when the output of an industry is expanded—the cost—is the value of what would have been produced elsewhere if that expansion had not taken place.

As the amount that has to be paid for resources is in general a reflection of the value of their output in other employments, in reckoning the cost of an additional supply, it will usually be good enough to take account of the value of the resources diverted to the production of that additional supply, or, alternatively, the value of the resources released if the output were not increased, without having to investigate what the resources would otherwise have been used to produce. On the other hand, and this needs to be borne in mind when you are dealing with the problems facing a commission, there may be cases in which the amount which you have to pay for factors does not represent the value of their output in an alternative use, and this they would therefore have to take into account.

In calculating the costs of an additional supply of a public utility service, it is of course necessary to start with the industry as it is, with whatever assets it possesses and the circumstances in which it finds itself. Costs are rooted in the actual situation. It is, for example, quite clear that the cost of an increase in the consumption of a particular public utility service equal to 5 percent of present consumption would be very different if the consumption of that service were declining by 8 percent per annum from what it would be if the consumption of that service were increasing by 8 percent per annum.

All this, I think, makes it clear that the costs of the supply of a public utility service are not the same thing as the monetary obligations which have been imposed on or assumed by the public utility concerned. The monetary obligations of the industry depend on all sorts of past decisions on financing and similar matters which have been made by the industry and also, presumably, the relevant commission. But the sum of money which has to be raised from consumers if the corporation is not to default on its obligations or is not to disappoint the expectations of its stockholders or management is one thing. The value of the resources that would be released if the supply of the public utility service, or part of it, were curtailed or stopped is quite another thing. To refer back to my previous example, a sudden change from a growth rate of 8 percent per annum to a decline of 8 percent per annum might have, at least immediately, a relatively small effect on the monetary obligations of the industry—the sum of money that has to be raised from consumers if there is not to be default or disappointed expectations—but the cost calculations would almost certainly be completely changed.

Of course, in deciding how much of the total sum of money to be raised by the industry is to be contributed by the various categories of consumers, and also the form of the rate structure, cost calculations have an important part to play. The cost of a resource represents its value in its best alternative use. The amount which consumers are willing to pay for it in a particular use represents its value in that use. It is only possible to be sure that resources are used in those ways which maximize the value of production if consumers are asked to pay an amount at least equal to the cost of the resources which are used on their behalf. If consumers are willing to pay an amount greater than the cost, then we know that the value of the product yielded by the resource is greater in this use than it would be in any other use. However, if consumers are to be required to meet this test, it is obviously important that costs should be calculated properly, and this means that they should accurately measure the fall in the value of production elsewhere if the particular supply of the public utility service under consideration is made available.

It is, I think, obvious that the so-called fixed costs are quite irrelevant in such a calculation. Items which remain the same whether or not a particular supply is undertaken cannot possibly be a cost of that supply. The cost is what you lose elsewhere by undertaking a supply. It is therefore only items which vary when that supply is undertaken or not undertaken which can come into a cost calculation. Similarly, so-called historical costs are also irrelevant. What will be lost if a supply is undertaken or continued is something which occurs in the present or the future. What may have happened in the past, or what expenditures have been made in the past, is not relevant to a cost calculation. A delving into historical records, except as a means of discovering what is likely to happen in the future, is a sure sign that the cost calculators are on the wrong track.

The sum of money which has to be raised from consumers to enable the utility to meet its financial obligations may require the price of some or all services of the utility to be raised above the cost of those services, but in setting the structure of prices so as to produce a given net revenue, it is desirable to do this in a way which minimizes the fall in the value of production, a way which minimizes the uneconomic diversion of resources.

An illustration: the TELPAK service

■ I propose to illustrate the role which costs should play in the determination of prices by examining the "Above 890 Megacycle" decision of the FCC, which resulted in the introduction of the TELPAK service by AT&T, and the policy discussions to which these events led. The setting within which this question must be examined arises out of a decision by the FCC over the allocation of frequencies in the bands over 890 megacycles, as a result of which the regulated common carriers in the communications industry became subject to potential competition from the establishment of private microwave systems on a greatly expanded scale.

I do not wish to quarrel with this initial decision. Indeed, I welcome it. The decision of the FCC involved an explicit recognition of the valuable role which the forces of competition can play in promoting the development of the communications industry. The competitive struggle in business, as elsewhere, is a stimulus to improvements of all kinds. One business introduces a better or a cheaper product, and others are forced to do the same. Indeed, the existence of competition will lead each business independently to make improvements, since a firm which merely matches the advance made by others runs the risk of being left behind. The contribution of this competitive process to the development of the American economic system is generally understood, and it is good to see that the FCC recognizes it.

The decision of the FCC in the "Above 890 Megacycle" case exposes the common carriers to much more extensive competition or potential competition in the microwave field from suppliers whose pricing and business policies are not subject to detailed control by any administrative agency. It is not my intention to argue for regulation of these new competitors of the common carriers. The way to solve the problems of competition is certainly not to regulate it out of existence. There are few more unpleasant sights than an unholy alliance between the regulators and the regulated industry to solve the problem of competition by suppressing it.

Let us rather welcome competition but face the issues it poses. The question then is to determine the proper response of the common carriers to the new situation. Are the existing prices and other policies of the common carriers in the communications industry appropriate in a regime of competition?

The new choice which opened up with the "Above 890" decision is whether a business should install a private microwave system to meet its communications needs, or whether it should make use of the facilities provided by the common carriers. Naturally, in making such a decision a business will compare the costs it would incur if it constructed its own system with what it would have to pay the common carrier for the services which the private microwave system might be expected to render. If, for example, the annual costs of a given microwave system are expected to be $125,000 per annum, and the charge made by the common carrier for similar services was $150,000 per annum, the business would presumably decide to install its own private microwave system. But would this decision be in the public interest?

The answer to this question depends on the cost which the common carriers would incur if they provided these communication services to the business. The amount that it would cost the common carriers to provide these services to the business would be the amount by which their total costs would increase if they provided the services as compared with what the total costs would be if they did not. It is the additional costs which providing this service would impose on their system which are relevant.

If, as a result of providing this service, the total cost of the common carriers increased by $100,000 per annum, it seems clear that it would be in the public interest that the business should satisfy its communications needs by taking them from the common carrier rather than by constructing its own private microwave system. The costs of using a private microwave system to supply these services would be $125,000 per annum. The costs incurred if the common carrier supplied these services would be $100,000 per annum.

In this example the use by the business of the common carrier's facilities would be in the public interest, but if the regulatory commission, which in this case is the FCC, insists on the common carrier charging $150,000 per annum, the common carrier would not provide the communication services, the private microwave system would be built, and the best use would not be made of the nation's resources.

If, on the other hand, the commission allows a charge to be made of $120,000, private microwave systems would not be built. A better use would be made of the nation's resources. The burden on the rest of the economy would be reduced by $25,000. This is what the saving in cost means. Incidentally, of course, the net revenues of the common carrier would increase by $20,000 per annum.

Of course this is a very simplified account of a very complicated situation, but I think it gives the essence of what is involved. When I speak of the costs of the common carrier providing the service which would otherwise be provided by private microwave systems, it is the additional costs imposed on the carrier which should be calculated. It is unnecessary to consider what costs were in the past, only what they will be in the future. Nor is it necessary to make any allocation of joint or common costs.

It may be objected that this is not the usual procedure of the regulatory commission in calculating cost, and this is no doubt true. It is certainly the case that when AT&T, after the "Above 890" decision, introduced the TELPAK service, which reduced rates to users of bulk communications (those most likely to build their own microwave systems) and which in concept corresponded to what I have described as desirable, the reaction of the FCC could not be described as one of welcoming this initiative on the part of the company. Frosty hostility would be a better descriptive phase. In fact, I have not seen any evidence that the FCC really understood how these cost calculations ought to be made, and the fact that it ordered the seven-way cost study to be made, with its allocations of the unallocable, is positive evidence that it did not understand how one should make cost calculations. This does not mean that AT&T's own cost calculations were necessarily correct. But a critical examination of these calculations should have concentrated on whether they accurately measured the additional costs incurred by the system as a result of the introduction of the TELPAK service.

There is one objection to the TELPAK tariff which cannot, I think, be dismissed so easily, and that is the issue of discrimination. The reduction in rates, to put the matter in simple terms, could only be obtained by the large user who might otherwise build a private microwave system. It was therefore denied to the small user who was not in a position to build such a system. The tariff, therefore, gives a competitive advantage to the businesses which get this reduction in price, which does not correspond to differences in the costs of providing these services. It will therefore tend to lead to some misallocation of resources.

The alternative to such a selective price reduction would of course be a reduction in price to every user, but this is something which presumably AT&T would not have proposed nor the FCC have accepted. Such a step would have threatened the financial stability of AT&T.

The practical alternative to be considered was a selective price reduction or no change at all. Without such a price reduction there would tend to be an uneconomic diversion of business from the common carriers to private microwave systems; with a selective price reduction there would tend to be some uneconomic diversion of business from the small users to the large users.

To an economist the question becomes one of deciding which of these is the lesser evil. In actual fact there is some reason to suppose that the uneconomic diversion of business from the small to the large user would tend to be rather small. Without the TELPAK service those businesses who are able to construct their own microwave systems would secure advantages denied to those who would not. The only uneconomic diversion that we need consider is that which comes about because the TELPAK tariff is less than the private microwave costs that would otherwise have been incurred. It is of course in AT&T's own interest to make this difference as small as possible.

There is another reason for thinking that a selective price reduction to the large user is not likely to have any serious economic consequences. The great bulk of the users of the TELPAK service were not in competition with the small users of the telephone service. By far the biggest user is the government, and no one I fancy is likely to be harmed by a reduction in the cost of government services. Furthermore, the competitors of many of the commercial firms in a position to use the TELPAK service will also be firms that are able to use the TELPAK service. It will be one steel company as against another steel company, and this also weakens the view that there will be an uneconomic diversion. I do not say that it will not happen, but it reduces the importance of this distinction.

The proportion of the business, therefore, done under this tariff to which it is even relevant to raise the question of possible harmful discrimination is probably rather small. I could be persuaded that this is wrong, but this in any case is not my main point. I regard the question of discrimination as a genuine issue, but the way to decide whether a selective price reduction should or should not be approved is not to indulge in dreary and hair-splitting discussions about whether it is a like service (and there was a good deal of this in this particular case), but rather to consider whether on balance the use made of the nation's resources is likely to be better with or without the price reduction. This is a problem of economics rather than of semantics, and the way to tackle it is to decide what policy makes the best use of the national resources.

But the reason why it is essential for the common carriers to think in these terms in a situation such as gave rise to the TELPAK service is not this, although this approach will contribute to the efficient use of the national resources. The reason is, to put it bluntly, that a public utility company in such a situation has no alternative in a regime of competition if it is to survive. The competitors of the public utility are going to base their policies on calculations of additional costs, and these will be current and future costs. Unless a public utility is allowed to do the same, it will gradually lose ground. The object of regulation is not presumably to kill the industry regulated. In a competitive situation, I suggest that a public utility company should be allowed to compete. I regard the TELPAK service therefore as a proper response to a competitive situation.

I regard this example as a good illustration of the value of economists' way of looking at cost because the issues seem to me relatively clearcut. I can imagine cases in which the issues are not so clearcut and in which, therefore, we would have to actually undertake the task of deciding whether the misallocation in this direction is worse than the misallocation in some other direction. When we come to that task, we will not find ourselves as well equipped to handle it as we are to handle such a relatively straightforward case as this.

References

1. CLEMENS, E. W. "Price Discrimination in Decreasing Cost Industries," *American Economic Review*, Vol. 31 (1941), pp. 794–802.
2. COASE, R. H. "Price and Output Policy of State Enterprise: A Comment," *Economic Journal*, Vol. 55 (1945), pp. 112–13.
3. ———. "The Marginal Cost Controversy," *Economica*, Vol. 13 (1946), pp. 169–82.
4. ———. "The Marginal Cost Controversy: Some Further Comments," *Economica*, Vol. 14 (1947), pp. 150–53.
5. ———. "Comments" in *Business Concentration and Price Policy*, A Conference of the Universities—National Bureau Committee for Economic Research. Princeton, N. J.: Princeton University Press, 1955. Pp. 436–38.
6. FLEMING, J. M. "Price and Output Policy of State Enterprise: A Comment," *Economic Journal*, Vol. 54 (1944), pp. 328–37.
7. HOTELLING, H. "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, Vol. 6 (July 1938), pp. 242–69.
8. LERNER, A. *The Economics of Control*. New York: MacMillan Co., 1944.
9. LEWIS, W. A. "The Two-Part Tariff," *Economica*, Vol. 8 (1941), pp. 294–370.
10. MEADE, J. E. "Price and Output Policy of State Enterprise," *Economic Journal*, Vol. 54 (1944), pp. 321–28.
11. NORDIN, J. A. "The Marginal Cost Controversy: A Reply," *Economica*, Vol. 14 (1947), pp. 134–49.
12. NORRIS, H. "State Enterprise and Output Policy and the Problem of Cost Imputation," *Economica*, Vol. 14 (1947), pp. 54–62.
13. PAINE, C. L. "Some Aspects of Discrimination by Public Utilities," *Economica*, Vol. 4 (1937), pp. 425–39.
14. RUGGLES, N. "The Welfare Basis of the Marginal Cost Pricing Principle," *Review of Economic Studies*, Vol. 27 (1949–50), pp. 29–46.
15. ———. "Recent Developments in the Theory of Marginal Cost Pricing," *Review of Economic Studies*, Vol. 27 (1949–50), pp. 107–26.
16. THIRLBY, G. F. "The Marginal Cost Controversy: A Note on Mr. Coase's Model," *Economica*, Vol. 14 (1947), pp. 48–53.
17. VICKREY, W. "Some Objections to Marginal Cost Pricing," *Journal of Political Economy*, Vol. 56 (July 1948), pp. 218–38.
18. ———, "Some Implications of Marginal Cost Pricing for Public Utilities," *American Economic Review Papers and Proceedings*, Vol. 45 (May 1955), pp. 605–20.
19. WILSON, T. "Price and Output Policy of State Enterprise: A Comment," *Economic Journal*, Vol. 55 (1945), pp. 454–61.

Resource allocation and the regulated firm

Elizabeth E. Bailey and John C. Malone

Data Systems Engineering Division
Bell Telephone Laboratories

This paper extends a simple economic model of the regulated firm to include a wider range of both management objectives and of regulatory constraint. The authors find that these generalizations in the assumptions are able to effect profound alterations in the conclusions. For example, if a firm is constrained to a fair return on invested capital, we find that the profit-maximizing firm would have an incentive to overcapitalize (the Averch-Johnson result), while the sales or output maximizing firm would have an incentive to undercapitalize. If any of several other methods of constraint are used, the authors find that there is no incentive to alter the relative amounts of labor and capital employed in the production process.

■ Considerable attention has recently been focused upon economic models describing the impact of regulation on a monopoly firm.[1] The model developed in this paper considers a generalization of these earlier models, under which a more general treatment of management objectives and of regulatory constraint is introduced. Management is seen as concerned either with total profits, with return on capital investment, with dollar sales, or with output. We consider several possible choices of the regulatory constraint. These include the usual fair-return-on-investment method as well as methods under which the firm is limited to a fixed profit, a fair profit per unit of production, and a profit equal to a fair fraction of total costs. Our model is similar to its predecessors in that it is a "static" model with many simplifying assumptions; for example, the demand and production functions do not vary with time, taxes and depreciation are ignored, factor costs are assumed constant, and there is no regulatory lag.

The constrained systems are analyzed to determine the status of resource allocation under the various management objectives and regulatory constraints. We are thus able to specify under what conditions the firm would overcapitalize (select a technology which uses

---

Mrs. Bailey received the B.A in Economics from Radcliffe College and the M.S. in Mathematics from Stevens Institute of Technology. She has been with Bell Telephone Laboratories since 1960, and is currently working toward the doctoral degree in economics at Princeton University.

Mr. Malone received the B S in Electrical Engineering from Yale University in 1963, the M.S. in Industrial Engineering from Johns Hopkins University in 1965, and the Ph D in Operations Research from Johns Hopkins in 1969. He was employed by Bell Telephone Laboratories from 1963 to 1968 and is now an associate of McKinsey & Company.

Both authors are indebted to R. D Coleman of Bell Laboratories for his invaluable assistance during the preparation of this paper, they are particularly grateful for his formulation of the proof in section 2–3. They are also indebted to many others, both in the Bell System and in the academic community (particularly O. Eckstein and W. J. Baumol), who have read the manuscript at various stages Views expressed are those of the authors and do not reflect official opinion or policy either of the Bell System or of McKinsey & Co.

[1] The first such model was that of Averch and Johnson [1]. Similar ones have been advanced by S. H Wellicz [23], O. E. Williamson [25], F. M. Westfield [24], A. K. Klevorick [11], and M. Z. Kafoglis [8]. See also [5], [7], [9], [20], and [22]

more capital and less labor than are consistent with minimum-cost operation), when the firm would undercapitalize, and when the firm would select an optimal (minimum-cost) combination of resources. We also indicate when the firm has an incentive to increase its production costs by some arbitrary means, such as maintaining idle or nonproductive resources.

For the company which is constrained to a fair return on investment, we find an incentive to use a nonminimum cost allocation of resources no matter which assumption is made about management objectives. The profit-maximizing firm will overcapitalize (the well-known Averch-Johnson effect), while the maximizer of dollar sales or physical volume will undercapitalize. For the company which has a return-on-investment objective, we find an incentive to undercapitalize which is not counteracted by any of the methods of constraint. Indeed, return-on-investment, whether treated as an objective or as a constraint, appears to produce a nonoptimum allocation. A constraint based on a profit ceiling or on a fair mark-up on cost may also affect the resource allocation. In this case, it is not the allocation between capital and labor which is changed; rather, we find that if demand is inelastic, then the profit or sales maximizing firm may have an incentive to employ resources nonproductively.

1. Description of the model

■ **1. Management Objectives.** The inclusion of profit maximization as an objective follows the traditional lines of microeconomic theory. The consideration of other objectives has been fairly common in recent literature, and indeed some precedents exist in the more historical literature. For example, Keynes observed that when stockholders are "almost entirely dissociated from the management, . . . the direct personal interest of the latter in making a profit becomes quite secondary."[2] Recent empirical investigations have confirmed Keynes' observation that managers may pursue goals quite different from owners; these studies have found that the assumption of profit maximization is "considerably less applicable in the case where the firm is controlled by management."[3] Thus, objectives other than profits are reasonable for a large corporation even if it is not regulated; for a regulated firm, the constraint on profits does, at least in principle, allow the managers considerable flexibility in satisfying objectives other than profit maximization. The reason is, of course, that there are many methods of operation, and many prices and output levels, which could be set so that the constraint is exactly met.

The alternative objectives included in this paper are sales, output, and return on investment. These additional components are not meant to be exhaustive,[4] nor to portray accurately the total objective function held by the management of any real firm. Indeed, if such a function did exist, it would probably be very complicated, depending on the relative strengths of many interests and individuals. Rather, our intention is merely to select a reasonable set of components and determine what impact, if any, each has on the firm's allocation of resources.

---

[2] See [10], p. 316.

[3] See [15], p 442.

[4] A more complete discussion of alternative objectives is given in [12]; see also [25].

The sales objective has been given serious consideration by Baumol, who has observed that in the business community any "program which explicitly proposes any cut in sales volume, whatever the profit considerations, is likely to meet a cold reception."[5] Baumol analyzes sales maximization subject to the condition that profits are maintained at some satisfactory level, and concludes that such a system will in general result in a lower price and larger output than would be determined under profit maximization.

The output or scale of operations objective can be traced back to Marshall, who stated that in large corporations expansion of operations may tend to be favored and indeed may proceed quite uncritically.[6] A recent treatment has been given by Kafoglis [8], who argues that a regulated firm which maximizes physical volume may produce at a higher level than either a profit or a sales maximizer.

Managers pay a great deal of attention not only to earnings but also to earnings-per-share. If the number of shares is fixed, as is often the case for firms which have substantial retained earnings, then the objectives are equivalent. However, for firms having large capital requirements (which includes most of the regulated public utilities), the number of shares is not fixed. In this case, an addition to the number of shares may well result in a dilution in earnings-per-share, even if total profits are increased. Any such decline in earnings-per-share might be expected to depress the price, at least under the assumption that the market views the price-per-share as a nondecreasing function of earnings-per-share. Hence, it is reasonable to suppose that managers might attach positive utility to earnings-per-share, and it should be worth considering this objective to see how it influences resource allocation.

There are many alternative ways of formulating such an objective for the purposes of model building.[7] For simplicity, we use a return-on-investment objective in our analysis.[8] Our reasoning is as follows. Either there is some value of the debt-to-equity ratio that minimizes the cost of capital or there is not. If there is, we assume the firm has found it, and we then observe that the same decisions are made by the firm if it maximizes return on investment as if it maximizes return on equity. If there is not, then the firm does at least select a long-run average proportion of debt to equity in its program of financing.[9] We then use this long-run average as the constant debt-to-equity ratio.

If we denote the firm's objective function by $X(L,K)$, the mathematical description of the problem is:

$$\underset{L,K}{\text{Maximize}}\ X(L,K)\,, \tag{1}$$

where L is labor and K is capital investment. We assume that $X(L,K)$ is one of the following four quantities:

$$\pi = \text{profit} = R(Q) - C(L,K)\,, \tag{2}$$

---

[5] See [2], pp. 47–8

[6] See [13], pp. 317–18.

[7] See, for example, Machlup [12], p 23.

[8] However, we should bear in mind the fact that neither return-on-investment, sales, nor output can be the only concern of an unconstrained firm, in the former case, we must insist on a reasonable volume of investment, and, in the latter cases, on the firm breaking even in its operations.

[9] See, for example, Modigliani and Miller [14].

$$r = \text{return on investment} = (R(Q) - wL)/K, \tag{3}$$

$$R(Q) = \text{revenues or sales} = p(Q)Q(L,K), \tag{4}$$

$$Q(L,K) = \text{level of production or output}, \tag{5}$$

where[10]

$$\begin{gathered} L \geq 0, \quad K \geq 0, \quad Q_L > 0, \quad Q_K > 0, \\ Q(0,K) = 0, \quad Q(L,0) = 0, \end{gathered} \tag{6}$$

$$p(Q) = \text{price, with } p' < 0, \tag{7}$$

$$C(L,K) = \text{cost} = wL + iK, \tag{8}$$

with w and i as constant wage and interest rates. An immediate consequence of the above is that

$$\begin{gathered} \pi = (r - i)K, \\ r = i + \pi/K. \end{gathered} \tag{9}$$

Several aspects of the objective functions and the definitions associated with them should be emphasized. In particular, the model assumes a single product and price with demand and production functions invariant over time. Equation (4) makes clear that the model we treat is an "inverse demand" model, in which the firm controls the amounts of labor and capital it employs.[11] The cost function, equation (8), ignores both taxes and depreciation, treats each resource as a scalar rather than a vector, and assumes that the factor markets are perfectly competitive (implied by the constant factor prices i and w).

☐ **2. Regulatory Constraint.** If the firm is regulated by some constraint upon its profits, the model becomes:

$$\underset{L,K}{\text{Maximize}}\ X(L,K) \tag{1}$$

subject to

$$\pi = fF(L,K), \tag{10}$$

where

$F(L,K) =$ some measurable characteristic of firm

$f =$ "fair" return on $F(L,K)$.

Thus, regulation limits profits to a fixed return on the measured characteristic; for example, if the characteristic is capital investment, then f will be the difference between a fair rate of return and the market cost of capital. The fixed return is assumed to be greater than zero, $f > 0$, so that revenues will exceed costs.

Naturally there is in general no guarantee that the firm will actually achieve the level of profit permitted by regulation. Effective-

---

[10] We use the usual notation for derivatives

$$R' = \frac{dR}{dQ}, \quad p' = \frac{dp}{dQ}, \quad Q_L = \frac{\partial Q}{\partial L}, \quad R_L = Q_L R'.$$

[11] An equivalent model can be constructed which makes explicit the notion that the monopoly firm also controls the price of its product, e.g.,

$$\underset{p,L,K}{\text{maximize}}\ pD(p) - wL - iK$$

subject to

$$D(p) = Q(L,K).$$

ness of the constraint depends on the value of f: when f is large enough, the constraint will not be effective and $\pi < fF(L,K)$; when f is small enough, the constraint will be effective and the equality will hold. If we let f^0 be the largest value of f such that the equality holds, then for $f > f^0$, the firm is unconstrained, and for $f \leq f^0$, the firm is constrained. We wish to consider only the case of the constrained firm; thus, we assume $0 < f \leq f^0$, and in this range, we may confine our attention to equation (10). We can thus use Lagrangian techniques rather than Kuhn-Tucker conditions.[12]

Since the system (1) and (10) is presented without reference to time, the constraint must be continuously maintained; this means that the regulatory agency reacts instantaneously to any change in condition, i.e., that there is no regulatory lag.[13]

The regulatory strategy that is of greatest practical interest, because of its wide use today, is the "fair" rate of return technique.[14] Under this method the firm's earnings may not exceed a fixed percentage, s, of its invested capital. There are at least two views on this "fair" rate. The first is that s constitutes as close an approximation to the company's cost of capital i as legal procedures and testimony can make it. The second view is that the regulator sets the rate s to exceed the cost of capital i, and thus allows the firm some positive profits over and above its cost of capital. It is the latter formulation which has been given the most attention in the economic models developed so far, and thus we shall restrict our attention to it. In terms of the general constraint given by equation (10), we can substitute $F(L,K) = K$ and $f = s - i$, so that

$$\text{Fair Return on Investment: } fF(L,K) = (s - i)K. \qquad (11)$$

The constraint is not only interesting because of its previous appearances in the literature,[15] but also because it is representative of a class of constraints in which the two factors of production are not treated symmetrically. We shall see that such lack of symmetry may result in nonoptimal allocation between labor and capital.

The fair mark-up on cost constraint is also of practical interest, since many aspects of governmental control are based on some variant of this regulatory method. Profits under this method are limited to a fixed percentage of costs, so that the general constraint becomes $F(L,K) = C(L,K)$:

$$\text{Fair Mark-up on Cost: } fF(L,K) = f(wL + iK). \qquad (12)$$

This method has the advantage of being relatively straightforward to apply, since costs are measurable. We shall see that it has the disadvantage of discouraging efficient operation by the firm whenever demand is inelastic.

The fixed profit per unit production strategy is of more theoretical than practical interest. The reason is, of course, that it will be difficult to justify selection of any particular fair unit profit, f, and it will also be difficult to define an appropriate unit of output for a real world with many differentiated products and services. The appeal of this method is twofold. First, the method has the obvious feature of encouraging the firm to produce more output. Second, the method will not distort resource allocation, either in the relative use of the two factors of production or in the wider sense of using all resources productively. Under the method, $F(L,K) = Q(L,K)$, so that

---

[12] Indeed, as Takayama [20] has pointed out, Averch and Johnson derived the conditions for the inequality (assuming return-on-investment regulation). The only case they analyzed, however, was the case of the equality.

[13] A treatment which incorporates lag appears in [2].

[14] See [4], [6], [16], [17], [18], and [19].

[15] See [1], [11], and [20].

$$\text{Fair Profit per Unit:} \quad fF(L,K) = fQ(L,K). \tag{13}$$

For instance, a power company might be allowed some fraction of a cent per kilowatt hour.

The last method of regulation we consider is the imposition of a profit ceiling on the firm so that

$$\text{Fixed Profit:} \quad fF(L,K) = M, \tag{14}$$

where M is some stated number of dollars. This method is interesting because it gives the internal allocation of resources that would occur in the unconstrained case. There would, of course, be a practical difficulty in selecting M. For example, the value of M would presumably be different for each of the various regulated industries; such differences would also occur within any particular industry at different points in time.

2. Analysis of the model

■ **1. Lagrangian Analysis.** We will analyze each objective function (2)–(5) subject to each constraint (11)–(14) to determine whether or not the firm finds it advantageous to use a nonoptimal allocation of resources, i.e., to operate at a point on an isoquant at which costs are not minimized[16]; for example, resource allocation is not optimal at either Q_1+ or Q_1- in Figure 1, since the cost associated with these points is C_2, which is greater than C_1, the cost at Q_1^*.

The condition that optimal allocation occurs at points of tangency between isocost lines and isoquants is a geometric representation of the optimality condition:

$$\frac{QK}{QL} = -\frac{dL}{dK} = \frac{i}{w}. \tag{15}$$

In words, the ratio of the marginal products equals the slope of the isocost line, which in turn equals the ratio of the two factor costs. Undercapitalization (use of a low-investment technology) occurs at points such as Q_1-, where

$$-\frac{dL}{dK} > \frac{i}{w}, \tag{16}$$

while overcapitalization (use of a high-investment technology) occurs at points such as Q_1+, where

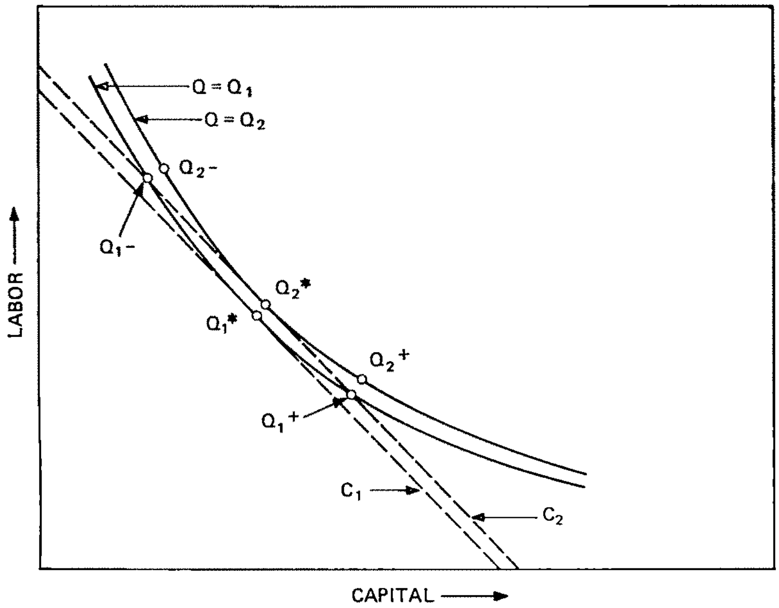
$$-\frac{dL}{dK} < \frac{i}{w}. \tag{17}$$

---

[16] A nonminimum-cost allocation of resources is often referred to in the literature as "misallocation"—see [1], [21], and [23]. We have deliberately avoided the use of this somewhat pejorative term, since it is not necessarily true that a nonminimum-cost allocation at some particular output level results in a loss of general welfare. For example, a firm would be overcapitalizing if it were to install a more expensive heating plant in an attempt to minimize air pollution. Such an action can be viewed as benefiting society by preventing external diseconomies of production, even though a nonminimum-cost allocation of resources has been used.

FIGURE 1

RESOURCE ALLOCATION



For the analysis of resource allocation, we form the Lagrangian function for each of the four objectives under consideration (π, r, R, and Q):

$$\pi\colon\ H(L,K,\lambda) = R(Q) - wL - iK - \lambda(R(Q) - wL - iK - [fF(L,K)]), \quad (18)$$

$$r\colon\ H(L,K,\lambda) = \frac{R(Q) - wL}{K} - \lambda\left(\frac{R(Q) - wL}{K} - \left[i + \frac{fF(L,K)}{K}\right]\right), \quad (19)$$

$$R\colon\ H(L,K,\lambda) = R(Q) - \lambda(R(Q) - [wL + iK + fF(L,K)'), \quad (20)$$

$$Q\colon\ H(L,K,\lambda) = Q - \lambda\left(Q - \left[\frac{wL + iK}{p(Q)} + \frac{fF(L,K)}{p(Q)}\right]\right), \quad (21)$$

where λ is the Lagrangian multiplier and $0 < f \leq f^0$. We observe that if $f > f^0$, then the firm's problem is to maximize only π or r or R or Q; thus, we may define $H(L,K,\lambda)$ for all $f > 0$ by taking $\lambda = 0$ when $f > f^0$, i.e., when the constraint is inoperative.

The four equations (18)–(21) have each been manipulated into the following general form:

$$H(L,K,\lambda) = X(L,K) - \lambda(X(L,K) - Y(L,K)), \quad (22)$$

where X is the objective function, and $(X - Y)$ is the constraint. Such a manipulation makes our analysis much simpler. We note that the first order conditions for a solution to (22) are

$$L\colon\ (1 - \lambda)X_L + \lambda Y_L = 0, \quad (23)$$

$$K\colon\ (1 - \lambda)X_K + \lambda Y_K = 0, \quad (24)$$

FIGURE 2

EVALUATION OF Y_L FOR LAGRANGIAN EXPRESSIONS

$H(L,K,\lambda) = X(L,K) - \lambda[X(L,K) - Y(L,K)]$

REGULATORY CONSTRAINT	MANAGEMENT OBJECTIVE			
	PROFIT	RETURN ON INVESTMENT	SALES	OUTPUT
FAIR % INVESTMENT	$0; Y_K = s - i$	0	w	$\frac{pw - (wL + sK)p'Q_L}{p^2}$
FAIR % COST	fw	fw/K	$(1+f)w$	$\frac{p(1+f)w - p'Q_L(1+f)(wL+iK)}{p^2}$
FAIR PROFIT PER UNIT	uQ_L	uQ_L/K	$w + uQ_L$	$\frac{pw + puQ_L - p'Q_L(wL + iK + uQ)}{p^2}$
FIXED PROFIT	0	$0; Y_K = -M/K^2$	w	$\frac{pw - (wL + iK + M)p'Q_L}{p^2}$

and

$$\lambda: \quad X - Y = 0. \tag{25}$$

If $\lambda = 1$, then interior solutions must have $Y_L = Y_K = 0$ for some L and K. Thus, if we can demonstrate by some outside argument that Y_L or $Y_K \neq 0$ for all L and K, then we can say that $\lambda \neq 1$. If λ is single-valued, continuous,[17] and never passes through 1, but is zero at $f = f^0$ (i.e., when the firm is not constrained), then we can conclude that $\lambda < 1$. Thus, we can guarantee that the sign of $(1 - \lambda)$ is positive.

Figure 2 shows the evaluation of Y_L for the four objectives (2)–(5) and four constraints (11)–(14). It is clear that in all but two of the cases either Y_L or $Y_K \neq 0$ for all L and K. By using the above argument, we can then conclude that $0 \leq \lambda < 1$ for these cases.[18] A separate analysis will be given for the two cases where the objective and constraint are identical.

The status of resource allocation for each particular expression is determined as follows. First, all terms containing Q_L or Q_K are retained on the left-hand side of equations (23) and (24), respectively, and all other terms are moved to the right. We then take the ratio of the resultant equations. For example, in the case of the output maximizer subject to rate-of-return regulation we find that (23)–(25) become

$$L: \ (1 - \lambda)Q_L + \lambda\left(\frac{pw - (wL + sK)p'Q_L}{p^2}\right) = 0, \tag{26}$$

$$K: \ (1 - \lambda)Q_K + \lambda\left(\frac{ps - (wL + sK)p'Q_K}{p^2}\right) = 0, \tag{27}$$

---

[17] Takayama points out that this assumption of continuity on λ may be a stronger one than appears on the surface and that in general it may require the continuity of the optimal L and K values with respect to f See [20], pp. 258–9

[18] Professor H. W. Kuhn of Princeton University pointed out to the authors that no genera argument limiting the range of λ can be made, hence we used a separate derivation for each case of interest.

FIGURE 3

RESOURCE ALLOCATION: RESULTANT RATIOS

REGULATORY CONSTRAINT	MANAGEMENT OBJECTIVE			
	PROFIT	RETURN ON INVESTMENT	SALES	OUTPUT
FAIR % INVESTMENT	$\frac{i}{w} - \frac{\lambda}{1-\lambda}\,\frac{s-i}{w}$	—	$\frac{s}{w}$	$\frac{s}{w}$
FAIR % COST	$\frac{i}{w}$	$\frac{r}{w} + \frac{\lambda f(r-wL/K)}{(1-\lambda-\lambda f)w}$	$\frac{i}{w}$	$\frac{i}{w}$
FAIR PROFIT PER UNIT	$\frac{i}{w}$	$\frac{r}{w} + \frac{\lambda}{1-\lambda}\,\frac{fQ/K}{w}$	$\frac{i}{w}$	$\frac{i}{w}$
FIXED PROFIT	—	$\frac{r}{w} + \frac{\lambda}{1-\lambda}\,\frac{M}{wK}$	$\frac{i}{w}$	$\frac{i}{w}$

and

$$\lambda:\ pQ - wL - sK = 0\,,$$

or, upon factoring Q_L and Q_K in (26) and (27),

$$Q_L\left(1 - \lambda - \frac{\lambda(wL + sK)p'}{p^2}\right) = -\frac{\lambda w}{p}\,, \tag{28}$$

and

$$Q_K\left(1 - \lambda - \frac{\lambda(wL + sK)p'}{p^2}\right) = -\frac{\lambda s}{p}\,, \tag{29}$$

whence, taking the ratio

$$\frac{Q_K}{Q_L} = -\frac{dL}{dK} = \frac{s}{w}\,.$$

Figure 3 summarizes the results of this type of analysis for each of the cases we consider. A discussion of these conclusions is given below.

☐ **2. Return-on-Investment Regulation.** For the profit maximizing firm subject to return-on-investment regulation, we see from Figure 3 that the resultant solution for resource allocation is exactly that derived by Averch and Johnson. Thus, their conclusion, stated as follows, holds.

> . . . if the rate of return allowed by the regulatory agency is greater than the cost of capital but is less than the rate of return that would be enjoyed by the firm were it free to maximize profit without regulatory constraint, then the firm will substitute capital for the other factor of production and operate at an output where cost is not minimized.[19]

A graphical interpretation of the Averch-Johnson result is given in Figure 4.[20] When profits are tied to the level of investment, then the greater the total investment, the greater the permitted profit. Since the volume of investments is under management control, management finds that its allocation of resources will affect the level of

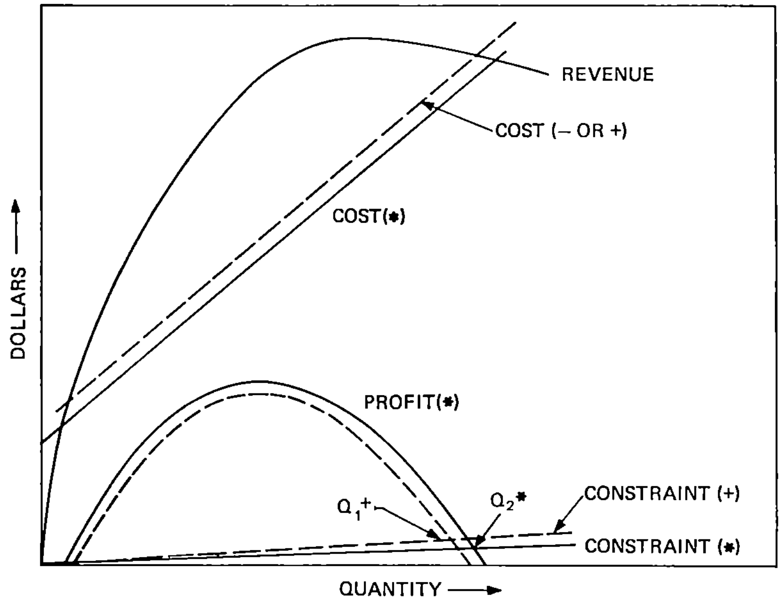
---

[19] [1], p 1053.

[20] A more sophisticated and quite elegant geometric explanation is given in [26].

FIGURE 4

RETURN-ON-INVESTMENT REGULATION



constrained profits. By overcapitalizing, management can increase allowed profits (dashed constraint in Figure 4). Of course, overcapitalizing increases costs as well (dashed cost curve in Figure 4), and management will not increase costs to such an extent that earned profits fall below allowed profits. In Figure 4 it is seen that the return-on-investment constraint can be satisfied at either Q_1+ or Q_2.* Of the two alternatives, the regulator clearly desires Q_2*, since it offers lower price and more of the product to the consumers. If the firm is strictly a profit maximizer, then management clearly desires Q_1+, since here the constrained profit is larger.

When the firm is a return-on-investment maximizer subject to a rate-of-return constraint, then the solution is no longer unique. Indeed, the solution is given by any pair of L and K that satisfy

$$\frac{R - wL}{K} = s\,. \tag{30}$$

Thus, the method of operation cannot be determined. If the firm is a sales or output maximizer, Figure 3 shows that inequality (16) holds; thus, the firm is shown to have an incentive to undercapitalize in its allocation of resources.

We have found that the impact of return-on-investment regulation on the firm's allocation of resources depends quite explicitly on our assumption about the firm's objective. If the objective is profit maximization, the constraint causes an incentive to overcapitalize. If the objective is output or sales maximization, the constraint gives the firm an incentive to undercapitalize. In the case where the objective and constraint are identical, the firm is indifferent among many possible methods of operation.

We could easily extend our results to the case where management has no single objective as such, but rather has a combination of various overlapping objectives. Consider, for example, the case of a firm interested in both profits and sales. If its objective were in the form

$$\underset{L,K}{\text{maximize}}\ \alpha\pi + (1-\alpha)R$$

then return-on-investment regulation would result in the following equation for resource allocation:

$$-\frac{dL}{dK} = \frac{i}{W} - \frac{\lambda}{\alpha - \lambda}\frac{(S-i)}{W}.$$

Thus, if $\lambda < \alpha$ the firm would overcapitalize, and for $\lambda > \alpha$, the firm would undercapitalize; i.e., the status of resource allocation would depend on the relative strength of the profit and sales objectives.

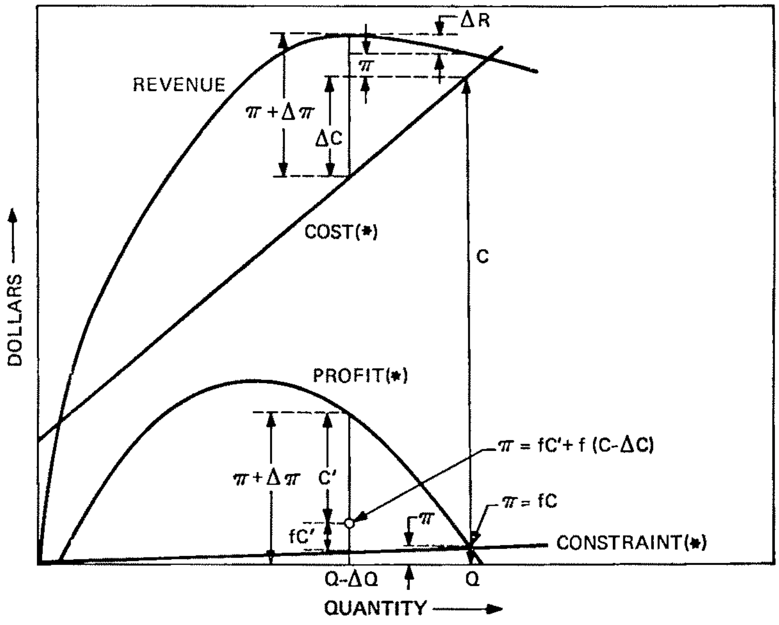
☐ **3. Constraint Based on Cost.** The implications for resource allocation under this method of regulation are derived after substituting $F_L = C_L = w$ and $F_K = C_K = i$ in (18)–(21). With a profit objective, Figure 3 shows that (15) holds and hence the firm has no incentive to distort the allocation of resources through either overcapitalization or undercapitalization. We will now show, however, that the firm may have an incentive to increase its costs by some arbitrary means.

Assume that the regulated firm is operating at Q in Figure 5, i.e., that it is using a minimum-cost production alternative and is satisfying the regulatory constraint. We then show that it is advantageous to the firm to increase costs by an amount C', since it can then increase its profits by fC' and still satisfy the regulatory constraint at some lower level of output $Q - \Delta Q$ (see Figure 5). So we compare the firm's profit at Q, defined as

$$\pi = fC \tag{31}$$

FIGURE 5

CONSTRAINT BASED ON COST



with the profit that it could achieve at the lower level of output $Q - \Delta Q$, which has a lower total cost, $C - \Delta C$

$$\pi + \Delta\pi > \pi > f(C - \Delta C). \tag{32}$$

In order for the constraint to be exactly met at the new level of output, the firm increases its costs by an amount C' on which constrained profits are fC':

$$\pi + \Delta\pi - C' = (C - \Delta C + C')f. \tag{33}$$

It will be profitable for the firm to move from Q to $Q - \Delta Q$, and thence to increase its costs, if and only if

$$\pi + \Delta\pi - C' > \pi, \tag{34}$$

i.e., the new level of constrained profits must exceed the original level of constrained profits. Rewriting (34)

$$\Delta\pi > C'. \tag{35}$$

Using (35) with C' given by (33) and π given by (31),

$$\Delta\pi > \frac{fC + \Delta\pi + f\Delta C - fC}{1 + f},$$

whence

$$\Delta\pi + f\Delta\pi > \Delta\pi + f\Delta C,$$

or

$$\Delta\pi > \Delta C. \tag{36}$$

Since

$$\Delta\pi = \Delta R + \Delta C, \tag{37}$$

then (36) becomes

$$\Delta R > 0. \tag{38}$$

Equation (38) states that the profit-maximizing firm will have an incentive to increase its costs so long as the constraint f intersects the profit curve in the inelastic range of demand, i.e., in the range where decreasing output increases revenues. If the constraint intersects the profit curve in the elastic range of demand, i.e., in the range where decreasing output decreases revenues, then (38) states that the firm will not find it profitable to increase its costs.

For the firm operating under a return-on-investment objective, Figure 3 shows that there is an incentive toward nonoptimal allocation. For the firm operating under a sales or output objective, we have

$$-\frac{dL}{dK} = \frac{i}{w},$$

i.e., the minimum-cost production result. As in the case of profit-maximization, however, the sales-maximizing firm will want to operate where marginal revenue is zero, even if it means increasing costs by some arbitrary means.

☐ **4. Fixed Profit per Unit of Production.** It is evident from Figure 3 that the desirability of this particular regulatory constraint depends on the firm's objectives. Minimum-cost production is employed under the profit, sales, and output objectives. However, for the return-on-investment case, profit-per-unit production regulation does not result in optimal allocation of resources; the constraint does not counteract the undercapitalizing tendency which is caused by the firm's desire to maximize the return on its investment.

Unlike the mark-up on cost constraint, however, fair profit per unit of production regulation does not cause any incentives for the firm to maintain idle resources. On the contrary, the constraint gives a strong incentive toward efficient operations.

☐ **5. Profit Ceiling.** We note that the profit-maximizing firm will be indifferent among any set of operating circumstances allowing a profit of M dollars. The sales maximizer will prefer to use a minimum-cost technology so long as demand is elastic. If, however, the constraint intersects the profit hill in the inelastic portion of the demand curve, the firm would have an incentive to cut production and maintain unproductive resources in order to increase its revenues. The output maximizer will use a minimum-cost technology no matter what the elasticity of demand. As in the other cases, the return-on-investment maximizer will have an incentive to undercapitalize.

3. Conclusions

■ The status of resource allocation in a regulated firm is extremely difficult to assess. Part of the difficulty is that the real world does not provide an unchanging production or market. Another aspect of the difficulty lies within the static analysis, for we have shown that the conclusions of economic models of the regulated firm are very sensitive to generalizations and alterations in the assumptions. We have only considered two areas for extension—management objectives and methods of constraint—and have found that even these have been able to effect profound alterations in the conclusions. For example, if a firm is constrained to a fair return on invested capital, we have found that the profit-maximizing firm would have an incentive to overcapitalize, while the sales or output maximizing firm would have an incentive to undercapitalize. For a management which is somewhat interested in both sales and profit, either conclusion might apply.

The various methods of constraint also differ in their impact on resource allocation. The fair-return-on-capital constraint was the only method which affected the relative amounts of labor and capital used in the production process. The fair mark-up on cost and the fixed profit constraints were both found to encourage nonproductive use of resources if demand were inelastic. We have also found only one constraint which would guarantee the use of minimum-cost operation by a firm which maximizes either sales or profits. This constraint, which is based on a fair-profit-per-unit output, not only allocates resources efficiently between labor and capital but also discourages any maintenance of nonproductive resources.

References

1. AVERCH, H. and JOHNSON, L. L. "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, Vol. 52 (December 1962), pp. 1053–1069.
2. BAILEY, E. E. and COLEMAN, R. D. "A Mathematical Model for Regulatory Lag," unpublished memorandum, Bell Telephone Laboratories, October 1969.

3. Baumol, W. J. *Business Behavior, Value and Growth.* New York: Harcourt, Brace & World, 1959.
4. Bonbright, J. C. *Principles of Public Utility Rates.* New York: Columbia University Press, 1961.
5. Bower, R. S. "Rising Capital Cost Versus Regulatory Restraint," *Public Utilities Fortnightly*, Vol. 75 (March 4, 1965), pp. 31–33.
6. Caves, R. E. "Direct Regulation and Market Performance in the American Economy," *American Economic Review*, Vol. 54 (May 1964), pp. 172–181.
7. Howard, R. A. and Matheson, J. E. "A Preface to a Theory of Regulation," Report No. 7379B–6, Stanford Research Institute, February 1969.
8. Kafoglis, M. Z. "Output of the Restrained Firm," *American Economic Review*, Vol. 59 (September 1969), pp. 583–9.
9. Kahn, A. E. "The Graduated Fair Return: Comment," *American Economic Review*, Vol. 58 (March 1968), pp. 170–73.
10. Keynes. J. M. *Essays in Persuasion.* London: 1931.
11. Klevorick, A. K. "The Graduated Fair Return: A Regulatory Proposal," *American Economic Review*, Vol. 56, (June 1966), pp. 477–84.
12. Machlup, F. "Theories of the Firm: Marginalist, Behavioral, Managerial," *American Economic Review*, Vol. 57 (March 1967), pp. 1–33.
13. Marshall, A. *Industry and Trade.* London: 1932.
14. Miller, M. H. and Modigliani, F. "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–1957," *American Economic Review*, Vol. 56 (June 1966), pp. 333–91.
15. Monsen, R. J., Chiu, J. S. and Cooley, D. E. "The Effect of Separation of Ownership and Control on the Performance of the Large Firm," *Quarterly Journal of Economics*, Vol. 82 (August 1968), pp. 435–51.
16. Phillips, A. and Williams, O. E., eds. *Prices: Issues in Theory, Practice and Public Policy.* Philadelphia, Pa.: 1968, pp. 108–123.
17. Rosoff, P. "The Application of Traditional Theory to a Regulated Firm," *Business Economics*, Vol. IV (January 1969), pp. 77–81.
18. Shepherd, W. G. "Regulatory Constraints and Public Utility Investment," *Land Economics*, Vol. 42 (August 1966), pp. 348–54.
19. ———, and Gies, T. G., ed's. *Utility Regulation, New Directions in Theory and Policy.* New York: Random House, 1966.
20. Takayama, A. "Behavior of the Firm Under Regulatory Constraint," *American Review*, Vol. 59 (June 1969), pp. 255–260.
21. Trebing, H. M., ed. *Performance Under Regulation.* East Lansing, Michigan: Michigan State University, 1968.
22. Weintraub, S. "Rate Making and an Incentive Rate of Return," *Public Utilities Fortnightly*, Vol. 81 (April 25, 1968), pp. 23–33.
23. Wellicz, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *Journal of Political Economy*, Vol. 71 (February 1963), pp. 30–43.
24. Westfield, F. M. "Regulation and Conspiracy," *American Economic Review*, Vol. 55 (June 1965), pp. 424–43.
25. Williamson, O. E. *Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm.* Englewood Cliffs, New Jersey: 1964.
26. Zajac, E. E. "A Simple Geometric Explanation of Averch and Johnson," *American Economic Review*, Vol. 60 (March 1970), pp. 117–125.

The valuation of public utility equities

Burton G. Malkiel

Professor of Economics
Princeton University

This paper reports on a cross-sectional valuation study of public utility equities during the year-end periods from 1961 through 1967. The ratios of market prices to earnings are related to such factors as anticipated earnings growth, dividend payout, and various proxy variables designed to measure the risk or quality of the earnings stream. The distinguishing feature of the study is that it uses the actual expectations of security analysts for variables that heretofore had to be estimated from historical data alone. The results of the regressions using expectations data are contrasted with results relying on historical data. Results of alternative risk proxies are compared, and the stability and predictive power of the model over time are examined.

1. Introduction

■ Much of the analytic financial literature of the past decade has assumed that the managers of firms are trying to make the shareholders as well off as possible. In many models this goal is stated in terms of maximizing the value of the common shares. If indeed managers do try to maximize this objective function in even some of their decisions, it is clearly essential that they know how their decisions affect the market prices of the shares. For example, if a firm's management is trying to decide on dividend policy, or on questions of financial structure, it must have some understanding of the effect of alternative possibilities on the price of the stock.

The valuation of a firm's common shares is also critically important to any study of the firm's cost of capital. While no attempt is made in this paper to estimate the cost of equity capital directly, cost estimates can be calculated from the expectations data used in the current study. It is worth emphasizing that public utilities are heavily dependent on outside capital; they have been accounting for over 20 percent of all new equity issues.[1] In view of the importance of outside sources of funds, knowledge of valuation relationships becomes a matter of great interest not only to utility managements but

---

The author is Professor of Economics and Director of the Financial Research Center at Princeton University He holds the B A. degree from Harvard University and the M B.A. degree from the Harvard Business School. He received the Ph D degree from Princeton University in 1964 From 1958 to 1960 Professor Malkiel was an associate in the investment banking firm of Smith, Barney, & Company in New York City. He is presently engaged in research on stock valuation, earnings forecasting, and the structure of the capital markets.

This paper was presented to the Conference on Public Utility Regulation sponsored by the AT&T Company and held at Stanford University on June 20, 1969 The empirical work is based on techniques developed jointly with John G. Cragg in a study of the valuation of corporate securities The author acknowledges helpful comments made by the editors of the Bell Journal of Economics and Management Science and by Myron Scholes He is also indebted to Jeffrey Balash, Raymond Hartman, Jane Ingersoll, and James Mann for their assistance in collecting and processing the expectations data, for programming assistance, and for carrying out the computations. This research was supported by the Institute for Quantitative Research in Finance and by the National Science Foundation.

[1] See [19], p. 17.

also to their investment bankers and, of course, to those charged with the responsibility of regulating public utilities.

Finally, a study of the valuation of public utility equities should be of considerable interest to investors concerned with knowing the relationship between stock prices and variables such as growth, payout, and risk. It is also important to know whether such relationships are stable through time or whether they change with market conditions, especially if one wishes to use empirical valuation models as a method of selecting "underpriced" stocks that might be expected to out-perform the market. Valuation models have been employed by Whitbeck and Kisor [27] and Peck [22] to select securities by comparing actual market prices with those predicted by cross-sectional regression analysis. It was assumed that the stock would seek its "warranted price" (the price indicated by the regression analysis) faster than that warranted price might itself change.[2] In the concluding section of this paper the results of such experiments with the present model are examined.

2. A model for the valuation of public utility equities

■ The typical stock-valuation model indicates that the present value of an equity share to the investor is the stream of returns to be expected from the security. In the simplest model, this is the present value of the stream of dividends, which is assumed to grow at a constant rate, g, over time.[3] Letting D stand for the (annual) dividend per share in the year just past and ρ the appropriate rate of discount, we have the present value of a share

$$P = D \sum_{t=1}^{\infty} (1+g)^t/(1+\rho)^t = D(1+g)/(\rho - g), \qquad (1)$$

provided $g < \rho$. Dividing by current earnings, E, we obtain

$$P/E = \frac{D(1+g)}{E(\rho - g)}. \qquad (2)$$

The price-earnings ratio is seen to depend on the long-term rate of growth and the dividend-payout ratio.[4]

The model presented above has several drawbacks. It cannot be employed when no dividends are currently paid, it leads to an infinite value for the shares when $g > \rho$, and it requires projecting a constant growth rate over an infinitely long horizon.[5] Such difficulties have led to the formulation of a finite-horizon model of share prices.[6] The basic idea of the finite-horizon approach is that dividends and earnings are assumed to grow at some rate g for T periods, and then to grow at some normal rate such as the general growth rate of the utility industry as a whole.[7] The present value of the estimated terminal price of the shares (at the end of T periods), plus the sum of the present values of the stream of dividends received through the T^{th} period, is taken to be the present worth of the shares. This approach can be illustrated by the following very simple model,[8]

---

[2] See [27], p. 344.

[3] See, for example, Williams [22] for one of the earliest statements of the problem and Gordon [8] for a more recent treatment

[4] The price-earnings ratio will be used as the dependent variable in the empirical work that follows, both because it seems more suitable when growth rates are a principal variable and because this specification helps reduce heteroscedasticity.

[5] Moreover, since the growth-rate estimates used in this study were specifically made for only the next five years, it would seem that this model is not consistent with the data.

[6] See, for example, Holt [10] and Malkiel [14]. Wendt [26] presents a useful survey of a number of the alternative models.

$$P_0 = \sum_{t=1}^{T} D_1 \frac{(1+g)^{t-1}}{(1+\rho)^t} + (m_s)_0 E_0 \frac{(1+g)^T}{(1+\rho)^T}, \tag{3}$$

where $(m_s)_0$ is the average current price-earnings ratio for the utility group and D_1 is the dividend expected to be paid next year. It will be noted that the formulation in equation (3) capitalizes the terminal year's earnings, $E_0(1 + g)^T$, at the average multiple for the utility group. This is consistent with the simplifying assumption that after the horizon period all utility stocks will enjoy (approximately) the same growth rate and thus revert to the same average condition.

Note that when $T = 1$ the P/E ratio for a share may be expressed as a linear combination of the growth rate and dividend-payout ratio

$$\frac{P_0}{E_0} = (m_s)_0\left[\frac{1}{1+\rho}\right] + \frac{D_1}{E_0}\left[\frac{1}{1+\rho}\right] + g\left[\frac{(m_s)_0}{1+\rho}\right]. \tag{4}$$

As T increases, the expression for P_0/E_0 becomes complicated with terms involving higher powers of the growth rate and dividend payout as well as cross-product terms in these variables. Fortunately, however, a linear approximation to the true expression seems to work reasonably well for T as small as five, the period for which growth-rate estimates are available.[9]

Thus far we have considered only the expected value of the stream of returns accruing to the shareholder. But a central feature of security evaluation is the expected uncertainty or instability (I) of these returns. Assuming that the typical investor is risk averse, he will prefer, other things equal, that this stream of returns be as stable and dependable as possible. Moreover, the horizon (T) over which extraordinary growth is forecast may itself be a function of the variance or dependability of the returns stream. Security buyers are more likely to project extraordinary earnings growth over a long horizon if the anticipated variance of earnings is low.[10] Since $\partial(P/E)/\partial T > 0$ for a growth stock according to the finite-horizon model developed above,[11] it follows that price-earnings multiples should be negatively related to a term estimating the expected variance of the earnings stream.[12]

---

[7] In some models the growth rate is assumed to decline in stages to the final "mature" growth rate of the economy In other models the initial and terminal growth rates are estimated on the basis of such factors as the retention rate and the rate of return on equity.

[8] The rationale for this approach and the derivation of equation (3) is contained in [14].

[9] The closeness of the proposed linear approximation was examined by the following experiment. Postulating values for the parameters $(m_s)_0$ and ρ that were consistent with experience during the 1961–1967 period, a series of theoretical price-earnings ratios for alternative pairs of payout rates and growth rates were calculated by equation (3). The theoretical P_0/E_0 ratios were then regressed on the alternative assumed values for payout and growth The coefficient of determination was over 0.97 When the square of the growth rate and the cross-product term $(D_1/E_0)g$ were included, the R^2 rose to approximately 1.00.

[10] See Peck [22].

[11] See Malkiel [14], pp. 1028–9.

[12] A more sophisticated argument for the inclusion of a variance term rests on recent theoretical work by Sharpe [24], Lintner [12], Mossin [20], and Malkiel and Cragg [15] extending the Markowitz [16] portfolio selection model. Under certain stylized assumptions the market es-

The preceding argument has focused on individual securities and not on each security's contribution to a total portfolio. A central feature of the Markowitz [16] portfolio model is that an "optimal" portfolio is not necessarily obtained by selecting those particular assets that are thought to be individually most desirable. To the extent that securities having low covariance can be found, the returns from a portfolio of such issues are likely to be more stable since a decline in the price of one security may be offset by independent behavior in the price of another. In Sharpe's [24] ingenious simplification of the Markowitz model, the returns from each security (R_i) are first related to the returns from some standard index of security prices, such as the Standard and Poor's composite average

$$R_i = \alpha_i + \beta_i(\text{Return From Index}) + \mu_i. \qquad (5)$$

Thus, covariances are assumed to arise because all returns depend on the common factor of the over-all market return. By analogy with the argument presented for the Markowitz model, we would expect investors to prefer those securities with low or negative β_i's. Other things being equal, a stock that tends to move against or only slightly with the market will tend to reduce the variability and, thus, the risk of the stock portfolio.

The final risk variable employed was a leverage variable. This is an index of the amount of debt in the firm's capital structure, i.e., the firm's "financial risk." The justification for its inclusion rests on the celebrated work of Modigliani and Miller [18] showing that, with perfect capital markets, the required rate of return on equity (which under certain simplifying assumptions is simply the inverse of the price-earnings ratio) will rise linearly with leverage. Even if one does not fully accept their analysis, however, earnings multiples still should be negatively related to leverage, since increasing leverage tends to increase the variance of the firm's earnings and also increases the firm's risk of ruin. Thus, while leverage should play no role in an equation where a fully adequate measure of risk is already included, leverage may well serve as a useful proxy both for the expected variability of the earnings stream and for the expected covariance with the market, β, as well.

This discussion may be summarized by outlining the following expected relationships. Price-earnings ratios should be positively related to expectations of future growth and dividend payout and negatively related to risk. The following alternative risk variables were employed: 1) the expected instability of the earnings stream, 2) an index of the conformance between an individual security's performance and that of the market, and 3) a leverage variable. In the following section we shall discuss the specific data employed in the study and indicate how they were collected.

3. Data sources and variables

■ The source of the principal data used in the study was a number of securities firms whose security analysts provided forecasts of the long-term growth rates of earnings, as of the seven year-end periods from 1961 through 1967, for a sample of electric utility common

---

tablishes "prices" for expected returns and for the variance of return associated with each security. Specifically, if we assume that the returns from different securities are uncorrelated with each other, it turns out that the price of a security should be a linear function of the expected return and the variance associated with the security.

stocks. Data were also collected on the analysts' estimates of "normal" earnings for the preceding year and their expectations about the future variability of earnings for the sample companies. Certain historical financial data were also used to provide a contrast with the expectations data. These included the past growth rate of earnings and three calculated risk proxies.[13] A detailed description of the data used in the study follows:

☐ **1. Normalized Earnings.** It is well known that the market does not necessarily capitalize the reported (after-tax) accounting earnings for a firm during the preceding year. Instead, investors employ some concept of normal earnings for each firm. Consider, for example, the hypothetical case of an electric utility that has been earning $5 per share for a number of years. Suppose that during 1969, as a result of damage caused by Hurricane Camille, earnings decline to $3 a share, the $2 difference being a nonrecurring loss. Will the market capitalize the $3 per share earnings? It is more likely that investors will recognize that this is a temporary dip in earnings and apply an appropriate price-earnings multiple to the amount they consider represents the normal earning power of the company. Indeed, one of the first jobs of a security analyst is to adjust the firm's accounting earnings to arrive at an indication of true earning power.[14]

The problem is particularly acute for utilities that market a large share of their output to cyclical industrial companies. The market does not apply a constant multiplier to cyclically varying earnings. Earnings multiples tend to fall as earnings rise, and rise as earnings fall. Thus, the price-earnings ratios that are relevant for valuation may be the ratios of prices to "normalized earnings" rather than ratios of prices to reported earnings. These normalized earnings are estimated to be the earnings that would obtain at a normal level of economic activity if the company were experiencing normal operations—that is, operations not affected by such nonrecurring items as strikes, natural disasters, and so forth.

There is an additional problem in using published accounting earnings figures as the basis of a valuation study. Reported earnings depend on a number of variable accounting procedures and management decisions. Consequently, the use of reported earnings figures would not be likely to put companies on a comparable basis. As Adam Smith puts it in the *The Money Game*,[15] there is the unique case of Zilch Consolidated, whose earnings can be played like a guitar by the use of different accounting treatments for depreciation, investment credits, inventories, and acquistions. While some of the more serious problems caused by acquisitions are largely absent in the case of utilities, many difficulties remain. Thus, one of the most important reasons for using normalized-earnings figures is to ensure that the same accounting conventions are applied to all companies. The convention applied to the normalized utility earnings figures used in the current study is that all data are adjusted by the analysts to allow tax savings from investments to "flow through" to earnings. The specific normalized-earnings figures used in the present study were averages of estimates supplied by the security analysts of two investment firms. The estimates were made by the analysts as of the end of each calendar year. Thus, estimates of, say, 1967 normalized earnings were made without any knowledge about subsequent earnings or even about the actual reported accounting earnings for 1967, which would not usually be reported until sometime early in 1968. While the estimates did differ somewhat between the two securities firms, care was taken to ensure that the same conventions were used by both firms in arriving at the normalized-earnings figures.[16]

---

[13] All historical data were taken from the COMPUSTAT tapes made available by Standard and Poor's Corporation.

[14] See Graham, Dodd, and Cottle [9], ch. 34.

[15] See [25], p. 182.

□ **2. Future Growth Rates.** As was mentioned above, several theoretical models of stock valuation have all focused on the expected growth rates of earnings and dividends as a central explanatory variable. Most previous empirical studies, however, were forced to rely entirely on published accounting data and past growth rates.[17] Whitbeck and Kisor [27] were able to increase the explanatory ability of their regressions by substituting the estimates of security analysts of one firm for fabricated values of expectational variables based on simple extrapolations of past performance. The present study substitutes the average estimates of earnings growth from several securities firms for the expectations of a single predictor.

Two types of growth rates were employed in the study, those collected from security analysts and those constructed entirely from historical data. The subjective long-term rates of growth were estimated by nine major securities firms.[18] Each growth rate figure was reported as an average annual rate of growth of (normal) earnings per share expected to occur over the next five years. The figures used in the study were the averages of the nine predictors. These expectations are not limited to published information. The security analysts involved frequently visit the companies they follow and discuss each company's prospects with its executives. Insofar as other security analysts follow the same sorts of procedures as our participating firms, the growth-rate estimates of other institutional investors and securities firms may resemble those employed in this study. Consequently, these predictions may well serve as acceptable proxies for market expectations.

In order to contrast the use of historical and expected growth rates, forty alternative historical growth rates were examined to find those that showed the closest correlation with market price-earnings multiples over each of the seven years covered by the study. These growth rates differed with respect to the period of calculation, the method of calculation, and the financial data upon which the calculation was made. From the forty candidates, one calculated growth rate was either clearly superior or, at least, no worse than any of the others in each of the seven years and was used in the regressions based on historical data. This was the ten-year growth rate of cash earnings

---

[16] Further information on the nature of the expectations data can be found in Cragg and Malkiel [5].

[17] For examples, see [2], [3], [4], [6], [7], [8], [10], [11], and [21].

[18] These data were mainly provided by the sponsoring members of the Institute for Quantitative Research in Finance. Five of the predictors were those described in Cragg and Malkiel [5]. All predictors were either securities firms or financial institutions that manage portfolios of common stocks. Growth-rate estimates were not available for the same sample of companies for each of the seven years included in the study. Consequently, sample sizes may vary from year to year.

TABLE 1
VARIABLES USED IN VALUATION STUDY OF PUBLIC UTILITIES

P	End-of-year market price per share.
D	Total dividends paid per share (adjusted to number of shares outstanding at year end).
E	Reported earnings per share (adjusted to exclude nonrecurring items).
$\overline{NE}$	Average "normalized" earnings estimates of security analysts.
$D/\overline{NE}$	Dividend-payout ratio.
$\bar{g}_p$	Average predicted future long-term growth rate of earnings per share.
g_H	Historical long-term growth rate of cash earnings per share.
I_p	Predicted instability index of the future earnings stream.
I_H	Historical instability index—semideviation of pre-tax earnings per share before fixed charges from trend.
β	The slope of a regression of the annual returns from a company's shares on the annual returns to the market index.
$F/(E+F)$	The ratio of fixed charges to earnings before fixed charges—a leverage variable.

per share (i.e., earnings plus depreciation and amortization) calculated as the geometric mean of first ratios.[19]

☐ **3. Dividend Payout.** The measurement of the dividend-payout ratio is not quite as straightforward as it might appear. If one simply takes the ratio of dividends to earnings, short-run disturbances to reported earnings that do not produce corresponding changes in dividends can make calculated payouts differ from target payout ratios. For this reason, payout ratios were calculated by dividing dividends by normal rather than reported earnings. In regressions using only historical data, the payout ratio was estimated from an average of the past seven years.

☐ **4. Risk Variables.** As mentioned above, three types of alternative risk measures were employed. The first was an instability index of earnings. An expected instability index was collected from one of the participating firms. It represented a measure of the past variability of earnings (around trend), adjusted by the security analyst to indicate potential future variability. A purely historical instability index was also employed. It was calculated as the semideviation of earnings before taxes and fixed charges around trend. As will be explained below, this variable is used as a proxy for the firm's "operating risk."

---

[19] A study, covering the 1962–1963 period, of the similarities between past growth rates and analysts' estimates of future growth, and of the accuracy of both sets of growth rates, is contained in Cragg and Malkiel [5].

TABLE 2

COMPARISON OF REGRESSIONS USING HISTORICAL AND EXPECTATIONAL VARIABLES

A REGRESSION RESULTS USING $P/\overline{NE}$ AS THE DEPENDENT VARIABLE											
$P/\overline{NE} = a_0 + a_1\bar{g}_p + a_2D/\overline{NE}$						$P/\overline{NE} = a_0 + a_1g_H + a_2D/E$					
YEAR	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	R^2	F	YEAR	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	R^2	F
1962	3.28	2.31 (0.27) 8 44	8 59 (5 49) 1 56	0 75	47 00 (2,31)	1962	4 85	1.64 (0 28) 5 87	10.86 (7.54) 1.44	0 61	24.55 (2,31)
1963	−4 80	2.99 (0 28) 10 67	14 79 (5 32) 2.78	0 81	67 39 (2,31)	1963	−4 58	2.11 (0.29) 7.40	19 93 (7.66) 2 60	0 68	33.55 (2,31)
1964	−2 58	2 97 (0.26) 11.48	12.58 (4 97) 2.53	0 84	80 77 (2,31)	1964	8.05	1 67 (0.35) 4.71	7 29 (9.27) .79	0.51	16.07 (2,31)
1965	3 02	2.45 (0 24) 10 23	5 01 (4.51) 1.11	0 84	86.60 (2,31)	1965	6.36	1 70 (0.29) 5.76	6 00 (7.37) .81	0 67	30.91 (2,31)
B REGRESSION RESULTS USING P/E AS THE DEPENDENT VARIABLE											
$P/E = a_0 + a_1\bar{g}_p + a_2D/\overline{NE}$						$P/E = a_0 + a_1g_H + a_2D/E$					
YEAR	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	R^2	F	YEAR	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	R^2	F
1962	2.59	2 39 (0 28) 8 46	8 62 (5.64) 1 53	0 75	47 51 (2,31)	1962	6.40	1.58 (0 31) 5 12	8.76 (8 31) 1 05	0.56	19.65 (2,31)
1963	−1.02	2 78 (0 24) 11 83	10 62 (4 47) 2 38	0 85	87 83 (2,31)	1963	0 56	1.89 (0 27) 7.06	14.07 (7.20) 1.95	0.68	33.24 (2,31)
1964	−3 91	3 10 (0.24) 13 02	12.81 (4 55) 2 81	0 87	104.65 (2,31)	1964	9 24	1 62 (0.37) 4 37	5.30 (9 75) .54	0 48	14 54 (2,31)
1965	−2 42	2 91 (0 26) 11.16	8.82 (4 92) 1 79	0.86	92 14 (2,31)	1965	3 38	1 93 (0.36) 5.42	8.17 (8 88) .92	0.63	26.17 (2,31)
C. LINEAR APPROXIMATION OF THEORETICAL RELATIONSHIP											
$P/\overline{NE} = -3.94 + \underset{(0.02)}{2.22}\bar{g}_p + \underset{(0.62)}{12.11}D/\overline{NE}$						$R^2 = 0.97$					

N.B. STANDARD ERRORS ARE RECORDED DIRECTLY BELOW THE COEFFICIENTS
WHILE "t VALUES" ARE RECORDED ON THE FOLLOWING LINE.
NUMBERS BELOW THE "F VALUES" ARE DEGREES OF FREEDOM.

The second type of risk measure introduced was the historical conformance between returns on each individual security and a market index. The variable employed was an estimate of the slope of a regression, where the dependent variable was the annual return from the i^{th} security and the independent variable was the annual return from Standard and Poor's 500-stock Average. Ten years of data were used in the calculation.

Finally, the leverage variable employed was the ratio of fixed charges per share to earnings, plus fixed charges per share. This variable, rather than the debt-equity ratio, was used to avoid the simultaneity problem of having the stock price depend on the leverage ratio whereas the debt-equity ratio by definition depends on the stock price.[20] Table 1 summarizes the variables used in the valuation study.

---

[20] For a discussion of the problems involved in using the debt-equity ratio itself, see Barges [1] and Wippern [29].

4. Analysis of the regression results

■ In this section the results of the basic certainty model (i.e., taking no account of risk) are first presented. Next, a comparison is made of the regression results for equations including comparable historical and expectational variables. Then, regressions employing the alternative risk variables are compared. Finally, results for the most satisfactory expectational equations are presented and the behavior of the coefficients over time is examined.

☐ **1. Results of the Basic Certainty Model.** In the left-hand side of part A in Table 2, results of the basic certainty valuation model are presented. Normalized utility earnings multiples are regressed only on growth and dividend payout, and no risk variables are included. We note that from 1962 through 1965 about 80 percent of the variance in price-earnings multiples is explained by the regressions. In all years the growth-rate variable is highly significant. The payout ratio always has the expected sign, but it is significant in only two of the four years. Plots of the data displayed no evidence of a nonlinear relationship between $P/\overline{NE}$ and $\bar{g}_p$.

In part C of Table 2, the equation of the linear approximation of the theoretical relationship is presented for comparison.[21] It will be noted that for most years the empirical valuation relationships are roughly similar to the theoretical one. However, the average growth rate coefficient over the four years studied is 2.68—about 20 percent higher than the theoretical coefficient of 2.22. On the other hand, the average dividend coefficient of 10.24 is almost 20 percent below the theoretical coefficient of 12.11. It would appear that investors in public utility stocks value dividends somewhat less than would be expected from the (nontax) theoretical present value model, whereas growth is valued more highly. These results are consistent with present tax laws, whereby returns are valued more when they are received as capital gains than as dividend income.

☐ **2. Comparison of Regressions Using Historical and Expectational Variables.** In the right-hand side of Table 2 the results of regressions using only variables calculated from readily available historical data are presented for comparison with the regressions employing expectations data. On the right-hand side of part A the (normal) earnings multiple is regressed on the historical 10-year growth rate of cash earnings and the actual dividend-payout ratio (averaged over the preceding 7 years). In part B of Table 2 the same comparison is made, but in these cases the actual earnings multiple rather than the normalized multiple is used as the dependent variable.

It will be noted that the fits are much better using the expectational variables than with the historical ones. It should also be mentioned that better fits were obtained by using the average growth

---

[21] The theoretical relationship was estimated by fitting a linear approximation to the finite-horizon valuation model described in equation (3). Specifically, estimates of $\hat{a}$, $\hat{b}$, and $\hat{c}$ were obtained by fitting a regression of the form

$$\frac{D_{j0}}{E_{j0}} \sum_{T=1}^{5} \frac{(1+g_j)^T}{(1+r)^T} + (m_e)_0 \frac{(1+g_j)^5}{(1+r)^5} = a + bg_j + c\frac{D_{j0}}{E_{j0}}.$$

A standard price-earnings multiple, $(M_e)_0$, of 17 (roughly the average over the four years studied) was used to compute the theoretical earnings multiples The discount rate, r, was estimated to be 9 percent. (Assuming an average payout for utilities of 60 percent, this implies a growth rate for the earnings of the market average of approximately 5½ percent per annum. See Malkiel [14], pp. 1011–12, for an explanation of these relationships.) As was mentioned above (footnote 9) a linear approximation fits the theoretical relationship quite closely.

TABLE 3

COMPARISON OF RISK VARIABLES, 1962–1965

INSTABILITY INDEX						INDEX OF CONFORMANCE WITH MARKET						LEVERAGE INDEX					
$P/\overline{NE} = a_o + a_1\bar{g}_p + a_2D/\overline{NE} + a_3I_p$						$P/\overline{NE} = a_o + a_1\bar{g}_p + a_2D/\overline{NE} + a_3\beta$						$P/\overline{NE} = a_o + a_1\bar{g}_p + a_2D/\overline{NE} + a_3F/(E+F)$					
YEAR	$\hat{a}_o$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	R^2	YEAR	$\hat{a}_o$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	R^2	YEAR	$\hat{a}_o$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	R^2
1962	5.51	2 26 (0 28) 8.01	7 33 (5 71) 1.28	–0.03 (0 03) –0 88	0 76	1962	3.94	2.29 (0.29) 8 01	8.10 (5 79) 1.40	–0.71 (2 27) –0.31	0 75	1962	3 86	2.31 (0.28) 8.35	9.37 (5 65) 1.66	–3.75 (5.45) –0.69	0 76
1963	–4.31	2.99 (0 28) 10 58	15 14 (5.41) 2 80	–0.02 (0.03) –0 55	0 82	1963	–4 14	2 98 (0 29) 10.46	14 22 (5 74) 2 48	–0 68 (2 36) –0 29	0.81	1963	–2.56	3 01 (0 24) 12 40	17.88 (4 70) 3 81	–15 23 (4.53) –3.36	0.86
1964	–1.81	2.96 (0 26) 11 20	12 56 (5 03) 2 50	–0.01 (0 02) –0.57	0 84	1964	–2 83	2 98 (0.27) 11 19	12.79 (5.29) 2 42	0 27 (2 07) 0 13	0 84	1964	–1.23	3 04 (0 24) 12 73	14.66 (4 62) 3.18	–11 23 (4 24) –2 65	0 87
1965	3.95	2 44 (0 24) 10 11	5 03 (4.53) 1.11	–0 02 (0 02) –0 85	0 85	1965	3.91	2 41 (0 25) 9.71	4 58 (4 59) 1 00	–1 00 (1 46) –0 69	0 84	1965	3.42	2 49 (0 24) 10 51	6 48 (4.54) 1 43	–6 06 (4 11) –1 47	0 85

N.B STANDARD ERRORS ARE RECORDED DIRECTLY BELOW THE COEFFICIENTS WHILE "t VALUES" ARE RECORDED ON THE FOLLOWING LINE. NUMBERS BELOW THE "F VALUES" ARE DEGREES OF FREEDOM.

CORRELATIONS AMONG RISK VARIABLES—1963.

	F/(E+F)	β
β	0 02	1 00
I_p	–0 17	–0 19

rates of all predictors than by employing the forecasts of any single investment firm. This suggests that a reasonable proxy has been obtained for what might be considered the expectations of the "representative" investor. It should also be noted that there was not too much difference in the regressions when the actual earnings multiple was substituted for the normal earnings multiple as the dependent variable.

The strong correlation between price-earnings multiples and predicted growth rates leads one to question the line of causality. Do stocks with high expected growth rates tend to sell at high price-earnings multiples because investors actively bid up the shares of companies with favorable prospects? Or does the security analyst see a large price-earnings ratio in the market and decide from this that the firm in question must indeed be a "growth stock?" One is reminded of the brokerage firm that was recommending a particular security because of its generous dividend yield. When the price of the security subsequently rose so that its dividend yield fell below average, the firm continued to recommend the security but simply transferred it from its "yield" list to its "growth-stock" list. Then, in order to justify the company's identification as a growth stock, the security analyst raised his estimate of the firm's long-term earnings growth.

Since past growth rates are also closely correlated with earnings multiples, however, it seems clear that the growth record of a company does influence the earnings multiples. This is not to deny, however, that expected growth rates may not be affected to some extent by the earnings multiples themselves. Nevertheless, this should not interfere with the basic purpose of this paper. The point is that even if these growth rates are simply the security analysts' estimates

TABLE 4

REGRESSION RESULTS EMPLOYING A LEVERAGE VARIABLE AND AN OPERATING-RISK PROXY

$P/\overline{NE} = a_0 + a_1\bar{g}_p + a_2D/\overline{NE} + a_3F/(E+F) + a_4I_H$

YEAR	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	R^2	CORRELATION BETWEEN $F/(E+F)$ AND I_H
1962	4 18	2.32	10 00	−4 38	−0.20	0 76	−0 09
		(0 28)	(5.70)	(5.50)	(0.21)		
		8.37	1.75	−0.80	0.97		
1963	−2.23	2.99	18 30	−15.78	−0.12	0 87	−0.12
		(0 25)	(4.77)	(4 63)	(0.17)		
		12.20	3.84	−3.41	−0 74		
1964	−0.20	2.99	14.51	−11 89	−0.17	0 87	−0.16
		(0.24)	(4.62)	(4.30)	(0 17)		
		12.32	3.14	−2.76	−0 97		
1965	4.03	2 48	6.26	−6.54	−0.09	0 85	−0.28
		(0.24)	(4.63)	(4.31)	(0.20)		
		10.23	1.35	−1.52	−0 45		

N.B. STANDARD ERRORS ARE RECORDED DIRECTLY BELOW THE COEFFICIENTS WHILE "t VALUES" ARE RECORDED ON THE FOLLOWING LINE. NUMBERS BELOW THE "F VALUES" ARE DEGREES OF FREEDOM.

of what the growth rate ought to be to justify the price, the regression results can still be used to gain an understanding of the structure of utility share valuations at any given time.

☐ **3. Analysis of the Alternative Risk Variables.** Table 3 presents an analysis of the alternative risk variables in the four years for which data on each of the three risk variables were available. Each risk variable was added in turn to the basic certainty equation employing predicted growth and normal payout. It will be noted that the leverage variable seems to be the most satisfactory risk index for utility shares. It always has the right sign and is comfortably significant in two of the four years. On the other hand, neither of the other two risk variables was ever significant, and β, the index of conformance with the market, has the wrong sign in 1964.[22] The table also suggests that leverage may itself work better as an *ex ante* instability index than the predicted instability index itself. It is also worth noting that the correlations among the risk variables tended to be insignificant. A typical correlation matrix for 1963 is presented at the bottom of Table 3.

In fact, however, the leverage coefficient may be biased toward zero if leverage is negatively related to operating or business risk. For example, utilities with low operating risk (i.e., stable revenues and earnings) may be willing to incur more debt than utilities with volatile operating earnings. High leverage might then compensate for low business risk, and the risk to the shareholder may be no

---

[22] It turned out, however, that the measured β's for utility stocks tended to cluster around the same value. Malkiel and Cragg [15] found considerably more variation in the β's for industrial securities, and, as one might expect, β added much more to an explanation of the $P/\overline{NE}$ ratios for industrials than for utilities.

TABLE 5

REGRESSION RESULTS: 1961–1967

$P/\overline{NE} = a_0 + a_1\bar{g}_p + a_2 D/\overline{NE} + a_3 F/(E+F)$						
YEAR	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	R^2	F
1961	8 11	2.63	1.95	−0.17	0 79	30.64
		(0.33)	(7.58)	(8.39)		(3,25)
		8.07	0.26	−0.02		
1962	3.86	2.31	9.37	−3.75	0.76	30.88
		(0 28)	(5.65)	(5.45)		(3,30)
		8 35	1.66	−0.69		
1963	−2 56	3 01	17.88	−15.23	0.86	63.75
		(0.24)	(4.70)	(4.53)		(3,30)
		12 40	3.81	−3.36		
1964	−1.23	3.04	14.66	−11.23	0.87	66.67
		(0.24)	(4.62)	(4.24)		(3,30)
		12.73	3.18	−2.65		
1965	3.42	2.49	6.48	−6.06	0.85	57.77
		(0.24)	(4.54)	(4.11)		(3,30)
		10.51	1 43	−1.47		
1966	2.42	2 32	1.93	−3.79	0.76	59.84
		(0.23)	(3.37)	(3.18)		(3,57)
		10 09	0.57	−1.19		
1967	−1 70	2 08	4 51	0.76	0 70	44.12
		(0 20)	(3.13)	(2.74)		(3,57)
		10.30	1 44	0.28		

N.B. STANDARD ERRORS ARE RECORDED DIRECTLY BELOW THE COEFFICIENTS WHILE "t VALUES" ARE RECORDED ON THE FOLLOWING LINE. NUMBERS BELOW THE "F VALUES" ARE DEGREES OF FREEDOM.

CORRELATION MATRIX FOR INDEPENDENT VARIABLES—1963

	$\bar{g}_p$	$D/\overline{NE}$
$D/\overline{NE}$	−0.53	1.00
$F/(E+F)$	−0.20	0.34

greater than for a utility with low leverage but more unstable operating income. In order to test for the importance of this phenomenon, a term was added to the regressions employing the leverage variable to account for the instability of the utility's pre-tax operating earnings before fixed charges. This instability index was calculated from a regression of the log of these earnings on time. It represented the semideviation of these earnings from trend. As Table 4 indicates, leverage is negatively correlated with the operating-risk proxy. When operating risk is added to the regression equation, the leverage coefficient becomes slightly more negative (and more significant) as expected, but the effects are so small as to be insignificant. Moreover, the operating-risk proxy itself is never significant, although it has the correct sign in all years.

☐ **4. Regression Results: 1961–1967.** In Table 5 results are presented for the regressions employing predicted growth, normal payout, and leverage as independent variables. Data were available to estimate equations with these variables for the seven year-end periods from 1961 through 1967. At the bottom of the table a typical correlation matrix of independent variables is presented. The only pair of entries displaying any substantial correlation is the payout ratio and the predicted long-term growth rate. This is hardly surprising, since, *ceteris paribus*, the more of its earnings a firm retains the larger its earnings should be in the future. Indeed, in many valuation models a firm's growth rate is estimated as the product of the firm's retention rate and its rate of return on equity.

It will be noted that the fits are extremely close for cross-sectional work. The regression results suggest that if one were given the most recent expectations figures regarding earnings growth, and if figures were available for dividend payout and leverage, then these data—together with the current values of the estimated coefficients—could be used to make a reasonably close prediction of the price-earnings multiple of the utility. However, these estimates, each for a particular year, only throw light on relative prices of stocks and not directly on the actual price level of any stock. In essence, the result merely says that the higher the anticipated growth rate the higher the normalized earnings multiple for the stock will be. This is hardly surprising, but it tells us that at least to this extent the market acts as an efficient allocator of capital. If many utilities were seeking capital at the same time, then those with the best growth prospects could sell their shares at a higher price than firms with lower expected growth rates.

☐ **5. Changes in the Valuation Relationship Over Time.** It is of considerable interest to examine whether or not the coefficients of the valuation equations are the same in each year or whether they change. This is important to investors who wish to use such valuation equations in connection with assigned values of the independent variables to estimate the "intrinsic worth" (or future price) of a security. Constancy of the relationship is also important if a firm is to seek to follow policies that will maximize the value of its shares. On the other hand, there is no reason to suppose that the valuation coefficients will be stable over time. As Table 5 indicates, the coefficients vary considerably over time.

We can observe how the individual coefficients of the valuation equation change from year to year. It will be recalled that at the end of 1961 "growth stocks were in high favor," and it is not surprising to find the growth coefficient in 1961 to be among the three highest coefficients recorded during the period studied. During 1962, however, there was a considerable change in the structure of share prices, which was acknowledged in the financial press as the revaluation of growth stocks. This revaluation is reflected in the coefficient of the growth rate, which declined considerably from 1961 to 1962. On the other hand, it is interesting to note that the dividend coefficient, which was among the lowest coefficients recorded in 1961, rose sharply by the end of 1962 as investors tended to put much lower weight on dividend payout. Similarly, the leverage variable, which was almost ignored by the market in 1961, becomes substantially more negative during the ensuing years.

It is also interesting to observe the behavior of the dividend and leverage coefficients during the 1966 and 1967 stock markets. It will be noted that the dividend coefficient falls to very low levels in these two years. Investors for whom utilities served as a substitute for fixed-income securities might well have forsaken utility stocks in favor of much higher yielding bonds.

The estimated dividend coefficients require some further explanation as the results are only partly consistent with some widely held notions about the valuation of public utility stocks. Most people believe that dividends are valued more highly for utility stocks than for the market as a whole. There is some support for this conjecture. Comparing the dividend coefficients for the utility stocks with those of a sample of industrials studied by Malkiel and Cragg [15], one finds that the utility coefficients tend to be higher than those estimated for the industrial stocks as a whole, although the differences between the two relationships are not statistically significant.

Utilities are often classified as stocks suitable especially for "widows and orphans," people dependent on their investment for a steady stream of income. Indeed, many security analysts believe that deviations from a high payout will have a particularly deleterious effect on the price-earnings multiples of utility shares. Furthermore, it is argued that another class of large holders of utilities is composed of tax-exempt foundations and educational institutions who gain no tax advantage by obtaining their returns in the form of capital gains rather than dividends. In addition, a number of these institutions adopt a convention of restricting their expenditures to the "income" from their portfolios, where this income is defined to include only dividends (and interest) received. Neither realized nor unrealized capital gains are considered as spendable income. Thus, because a large number of utility holders specifically desire income through dividend payments and are not affected by income taxes, it would appear that dividends would indeed contribute great weight to share valuations in the utility industry. While it is reasonable that dividends are valued more highly for utilities than for industrials, it must still be emphasized that the dividend coefficients tend to be lower than those we would expect from a theoretical present-value model. As was noted above, this is consistent with the differential tax treatment of dividend and capital-gain income for taxable investors.

The behavior of the leverage coefficients during the 1966–1967 period is somewhat difficult to explain. The coefficient associated with the leverage index becomes less negative during this period of very high interest rates. In 1967 the coefficient is actually positive but not significant. One might argue that as interest rates rise utilities with high leverage ought to be penalized more than utilities with low leverage, particularly if their outstanding debt tends to have a short average maturity. This is so because such firms will be faced with the prospect of refunding outstanding low-coupon debt issues with bonds of considerably higher cost. On the other hand, the leverage index may be directly related to the maturity of a utility's outstanding debt. Since leverage is defined as the ratio of fixed charges to earnings plus fixed charges, a firm that put out most of its debt several years ago may have relatively low fixed charges because of the low coupons prevailing at the time the bonds were issued. Of course, it is precisely such a firm that will suffer most from high interest rates, since it will have to refund all of its low-coupon debt at considerably higher cost. This may provide some explanation for the actual behavior of the leverage coefficient.

5. The empirical valuation model and security selection

■ One of the most intriguing questions concerning empirical valuation models is whether they can be used to aid investors in security selection. It has been suggested that empirical valuation models can be employed for security selection in the following way: The estimated valuation equation shows us, at a moment in time, the average way in which variables such as growth, payout, and leverage influence market price-earnings multiples. Given the value of these variables applicable to any specific utility, we can compute an estimated price-earnings ratio based on the empirical valuation equation. The next step is to compare the actual market earnings multiple with that predicted by the valuation equation. If the actual earnings multiple is greater than the predicted multiple, we designate the security as temporarily "overpriced" and recommend sale. If the actual price-earnings multiple is less than the predicted multiple, we designate the security as temporarily "underpriced" and recommend purchase. Such a procedure was employed by Whitbeck and Kisor [27] and Peck [22] with considerable success. They claimed that an underpriced group of securities selected by the above procedure consistently outperformed an "overpriced" group.

Of course, even on *a priori* grounds, it is possible to think of many reasons why such a procedure would prove nugatory. For example, consider what would happen if high P/E (high growth rate) stocks were overpriced in one period. In such a case, the estimated growth-rate coefficient will be larger (by assumption) than is warranted. However, the recommended procedure will not indicate that high P/E stocks are overpriced. This is so because the predicted earnings multiples for these securities will themselves be higher than is warranted since they come from regressions where the dependent variables are the actual earnings multiples themselves. Moreover, the very high correlations achieved cast doubt on our ability to forecast future performance. Indeed, if we were able to explain the existing structure of prices completely (i.e., there was perfect correlation), we would have no basis at all for selecting underpriced securities. Nevertheless, in view of the positive results reported by Whitbeck and Kisor and Peck it would seem desirable to attempt to replicate their experiment with the expectations data employed in this study.

The results of some of the experiments for the years 1962–1965 are shown in Table 6. The results are based on regressions shown in Table 5. The degree of over- or underpricing was taken to be the residual from the prediction equation, i.e., $(P/\overline{NE} - \hat{P}/\overline{NE})$. If the model is useful in measuring underpricing, then underpriced securities, according to this criterion, ought to "outperform" overpriced securities over some subsequent period. One year was chosen as the appropriate horizon, and subsequent returns were measured in the normal manner:

$$R_{t+1} = \frac{P_{t+1} - P_t + D_{t+1}}{P_t}. \qquad (6)$$

If the empirical valuation model is successful in selecting securities for purchase, the residual (i.e., the difference between the actual and predicted multiples, or the degree of overpricing) from the valuation equation ought to be negatively related to these subsequent returns. As Table 6 indicates, the correlations are very low and are essentially

TABLE 6

USE OF MODEL AS STOCK SELECTOR

1962–1965*

$R_{t+1} = b_0 + b_1 \left[P/NE - \widehat{P/NE}\right]$			
YEAR	$\hat{b}_1$	R^2	F
1962	0.50 (0.59) 0.84	0.02	0.71 (1,32)
1963	−0.10 (0.78) −0.13	0.00	0.02 (1,32)
1964	0.49 (0.78) 0.63	0.01	0.40 (1,32)
1965	1.01 (0.78) 1.30	0.05	1.68 (1,32)

*BASED ON REGRESSION RESULTS IN TABLE 5

N.B. NUMBERS IN PARENTHESES BELOW COEFFICIENTS ARE STANDARD ERRORS; NUMBERS BELOW PARENTHESES ARE "t VALUES", NUMBERS BELOW THE "F VALUES" ARE DEGREES OF FREEDOM.

zero. Similar results were obtained using the regression equations shown in Tables 2 and 3. The residuals from the equations employing historical data were no more successful in predicting subsequent performance. Moreover, other methods of measuring success in security selection such as examining the subsequent performance of, say, the five most overpriced and underpriced issues, were equally unsuccessful.

It is interesting to inquire why the model has been so unsuccessful in selecting securities with above-average subsequent returns. There appear to be three reasons for the lack of forecasting success.

One reason for the predictive failures has already been mentioned. The coefficients of the valuation relationship are sufficiently volatile that they cannot be used to establish valuation norms. By the next year (the end of the horizon period) the norms of valuation may be significantly altered. What was cheap on the basis of 1961's relationship (when growth was highly valued and dividend payout was almost ignored) may no longer represent good value on the basis of the relationship existing in 1962, as was indeed the case. This argument suggests that more success might have been achieved with a shorter horizon period. Nevertheless, tests employing a three-month horizon period proved just as unsuccessful.

Another reason for our lack of success in selecting securities with above-average subsequent returns was the generally poor quality of the earnings forecasts used as inputs to our study. Such a finding is really quite surprising. One would have believed that utility earnings are much more predictable than those of industrial companies. In a study by Cragg and Malkiel [5], however, it was shown that these growth rate forecasts proved little better than simple extrapolations of past growth rates in anticipating future earnings. While it is true that the mean squared errors for the utility forecasts tended to be lower than those of industrials, it turned out that security analysts' rankings of estimated growth within the utility industry were little more useful than random rankings. Rank correlations of predicted and realized growth rates within the utility industry were actually somewhat worse than those calculated for other industries. Needless to say, the availability of better forecasts would have improved the results substantially. This factor may be responsible for the differences in the predictive success of alternative empirical valuation models. The models themselves are probably much more similar than the inputs used to estimate them.

A final reason why the model may fail to predict performance may be specification error. Undoubtedly, there were special features applicable to many individual utilities that were not captured in our independent variables. That this conjecture has some validity is attested by an examination of the residuals from the equations in Table 4 for the years 1961-1964. A comparison of actual and predicted price-earnings multiples indicated that 14 stocks (well over one third of the sample) were consistently overpriced or underpriced in each of the four years. More than two thirds of the stocks in the sample were overpriced or underpriced in three of the four years. Consequently, it cannot be said that all deviations of actual from predicted price-earnings ratios are simply manifestations of temporary price anomalies.

One possible explanation for this result is the omission of variables. In the case of utilities, the most important missing factor in the valuation relationship may well be differences in the regulatory climate surrounding the various companies, which lead to differences in perceived risk. The fits might well be improved by adding dummy variables in the regression to capture regulatory differences, many of which can be categorized along geographical lines. Some preliminary experiments with this approach, however, do not indicate that such a procedure can improve the predictive ability of the regressions. Thus, despite the success in utilizing expectations data for estimating a valuation equation which has far more explanatory ability than one based on historical information, it is still quite clear that the present model does not offer a reliable guide for selecting securities.

References

1. Barges, A. *The Effect of Capital Structure on the Cost of Capital.* Englewood Cliffs, N. J.: Prentice-Hall, 1965.
2. Benishay, H. "Variability in Earnings-Price Ratios of Corporate Equities," *American Economic Review*, Vol. 51, No. 1 (March 1961), pp. 81–94.
3. Bower, R. S. and Bower, D. H. "Risk and the Valuation of Common Stock," *Journal of Political Economy*, Vol. 77 (May-June 1969), pp. 349–62.
4. Cohen, K. J. and Pogue, J. A. "An Empirical Evaluation of Alternative Portfolio-Selection Models," *Journal of Business*, Vol. 40, No. 2 (April 1967), pp. 166–93.
5. Cragg, J. G. and Malkiel, B. G. "The Consensus and Accuracy of Some Predictions of the Growth of Corporate Earnings," *Journal of Finance*, March 1968, pp. 67–84.
6. Fisher, G. R. "Some Factors Influencing Share Prices," *Economic Journal*, Vol. 71 (March 1961), pp. 121–141.
7. Friend, I. and Puckett, M. "Dividends and Stock Prices," *American Economic Review*, Vol. 54, No. 5 (September 1964), pp. 656–82.

8. Gordon, M. J. *The Investment, Financing, and Valuation of the Corporation.* Homewood, Illinois: Richard D. Irwin, 1962.
9. Graham, B., Dodd, D. L. and Cottle, S. *Security Analysis, Principles and Technique* New York: McGraw Hill, 1962 (4th ed.).
10. Holt, C. C. "The Influence of Growth Duration on Share Prices," *Journal of Finance*, Vol. 17 (September 1962), pp. 464–75.
11. Jen, F. C. "Common Stock Valuation: An Empirical Investigation of the Finite Horizon Hypothesis," a paper presented to the meeting of the American Statistical Association on September 10, 1965.
12. Lintner, J. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, Vol. 47 (February 1965), pp. 13–37.
13. Little, I. M. D. "Higgledy Piggledy Growth," *Oxford Institute of Statistics Bulletin*, Vol. 24 (November 1962), pp. 387–412.
14. Malkiel, B. G. "Equity Yields, Growth, and the Structure of Share Prices," *American Economic Review*, Vol. 53, No. 5 (December 1963), pp. 1004–31.
15. ——— and Cragg, J. G. "Expectations and the Structure of Share Prices," forthcoming.
16. Markowitz, H. *Portfolio Selection: Efficient Diversification of Investments.* New York: John Wiley & Sons, 1959.
17. Miller, M. H. and Modigliani, F. "Dividend Policy, Growth and the Valuation of Shares," *Journal of Business*, Vol. 34 (October 1961), pp. 411–33.
18. Modigliani, F. and Miller, M. H. "The Cost of Capital, Corporation Finance and the Theory of Investment," *American Economic Review*, Vol. 48, No. 3 (June 1958), pp. 261–97.
19. Morrissey, F. F. and Lynn, R. K. "Postwar Financial Experience of the Electric Utility Industry, Part I," *Public Utilities Fortnightly*, January 31, 1963, pp. 15–22.
20. Mossin, J. "Equilibrium in a Capital Asset Market," *Econometrica*, Vol. 34 (October 1966), pp. 768–83.
21. Ortner, R. "An Estimate of the Time Horizon and Expected Yield for a Selected Group of Common Shares, 1935–1955," *International Economic Review*, Vol. 2 (May 1961) pp. 179–98.
22. Peck, L. G. "Evaluation of Electric Utility Stock Prices," unpublished manuscript, Goldman, Sachs & Company.
23. Pratt, J. W. "Risk Aversion in the Small and in the Large," *Econometrica*, Vol. 32 (January-April 1964), pp. 122–36.
24. Sharpe, W. F. "A Simplified Model for Portfolio Analysis," *Management Science*, Vol. 9 (January 1963), pp. 277–93.
25. Smith, A. *The Money Game.* New York: Random House, 1967.
26. Wendt, P. F. "Current Growth Stock Valuation Methods," *Financial Analysts Journal*, Vol. 21 (March-April 1965), pp. 91–105.
27. Whitbeck, V. and Kisor, M. "A New Tool in Investment Decision Making," *Financial Analysts Journal*, Vol. 19 (May-June 1963), reprinted in Fredrickson, E., *Frontiers of Investment Analysis*, Scranton, 1965.
28. Williams, J. B. *The Theory of Investment Value.* Cambridge, Mass.: Harvard University Press, 1938.
29. Wippern, R. F. "Financial Structure and the Value of the Firm," *Journal of Finance*, December 1966, pp. 615–33.

HE BELL JOURNAL OF ECONOMICS ND MANAGEMENT SCIENCE

Editor:
Paul W. MacAvoy, *Massachusetts Institute of Technology*

Associate Editors:
Harvey J. McMains, *American Telephone and Telegraph Company*
William F. Sharpe, *Stanford University*

Managing Editor:
Gray L. Bromleigh, Jr., *American Telephone and Telegraph Company*

UTUMN '70

.. 1. NO. 2.

Table of contents

162 **Input choices and rate-of-return regulation: an overview of the discussion**
WILLIAM J. BAUMOL AND ALVIN K. KLEVORICK

191 **Peak-load pricing of telephone calls**
STEPHEN C. LITTLECHILD

211 **Rate of return and business risk**
PAUL H. COOTNER AND DANIEL M. HOLLAND

227 **The economics of railroading: the beginning of cartelization and regulation**
ROBERT M. SPANN AND EDWARD W. ERICKSON

245 **Investment and rate of return for the regulated firm**
BLAINE E. DAVIS

271 **The effectiveness of the Federal Power Commission**
PAUL W. MacAVOY

304 **A mathematical formulation of the peak-load pricing problem**
ISRAEL PRESSMAN

327 **Regulation of motor freight transportation: a quantitative evaluation of policy**
JAMES SLOSS

Input choices and rate-of-return regulation: an overview of the discussion

William J. Baumol
Professor of Economics
Princeton University

and

Alvin K. Klevorick
Associate Professor of Economics
Yale University

This article reviews the substance of the literature stemming from the Averch-Johnson model of the firm under rate-of-return regulation. It examines a number of propositions, among them the following: (1) The profit-maximizing firm under rate-of-return regulation will tend to use a capital-labor ratio greater than that which minimizes cost for its output level; (2) The profit-maximizing firm under regulatory constraint will use a capital-labor ratio and produce an output greater than it would in the absence of regulation; (3) The closer the "fair rate of return" is to the true cost of capital, the greater the quantity of capital the firm will want to use; and (4) The sales (total revenue)-maximizing firm under regulatory constraint will use a labor-capital ratio greater than it would when unconstrained. The second of these propositions, which has been widely taken to have been proved by Averch and Johnson, is shown to be false. The paradoxical assertion in the third proposition is explained and its regulatory implications discussed. Two models involving regulatory lag are described along with their implications for policy. Finally, some evaluative comments are offered on the entire issue of what has come to be known as the "A-J effect" and its importance in regulatory economics.

■ The imposition of an artificial constraint on a decision process that would otherwise be optimal will certainly lead to no improvement in outcome and will generally make things worse. In particular, a regulatory constraint on the firm's rate of return, if it affects input proportions at all, may be expected to reduce their efficiency. This issue was explored explicitly by Harvey Averch and L. L. Johnson

---

Dr. Baumol has taught at Princeton University since 1949. He was appointed Professor in 1954. He holds the B S.S. degree from the City College of New York (1942) and the Ph.D. degree from the University of London (1949). Dr. Baumol has published extensively on subjects ranging from the theory of the firm to the economics of the performing arts and has frequently been involved in proceedings relating to public utility regulation.

Dr. Klevorick has been with Yale University since 1967, where he is Associate Professor and a member of the Cowles Foundation. He was a Lecturer in Economics at Princeton in 1966–67. He holds the B. A. degree from Amherst College (1963), and the M. A. and Ph.D. degrees from Princeton (1965 and 1967, respectively). Dr. Klevorick is currently on leave-of-absence at the London School of Economics

The authors are grateful to the National Science Foundation and the Ford Foundation for support while this paper was being completed. Dr. Klevorick held an Air Force Office of Scientific Research–National Research Council fellowship which also facilitated this work. The authors are also grateful to Mrs. E. E. Bailey for invaluable criticisms and bibliographical assistance and to Dr. E. E. Zajac for helpful comments.

in a pathbreaking article in 1962 [1], using a model similar to one constructed almost simultaneously in a paper written independently by S. H. Wellisz[1] [11]. These authors showed within a static equilibrium framework that, given a number of conditions to be discussed presently, rate-of-return regulation will indeed modify the capital-labor ratio utilized by the profit-maximizing firm. Specifically, their model led to a demonstration that under the conditions postulated the firm will find it profitable to employ more capital relative to labor than is consistent with minimization of cost for the quantity of output produced.

These results have led to a considerable discussion, technical and informal, in terms of abstract general principles and concrete policy implications. The writings on the subject have disputed the rigor of the original proofs and supplied alternative derivations; they have raised questions about the assumptions and expressed doubts about the applicability or the importance of the results in practice.

Without surveying the literature item by item, this paper seeks to review the substance of the discussion. It first restates the relevant theorems and assertions and provides informal and formal proofs. It then turns to some related propositions that have been developed and discusses some work currently in progress in the area. Finally, it offers, for whatever they may be worth, our views of the upshot of the discussion. In sum, we conclude that the analytical approach employed by the original authors is a very fruitful one. Indeed, their piece represents a breakthrough in the theory of regulation, constituting a major contribution to the rigorous analysis of the principles of regulation of industry, a field which has recently blossomed. Moreover we agree that there are dangers to economic efficiency in the rate-of-return approach to rate regulation. Nevertheless, we will suggest that the phenomenon that emerges from the A-J theorem may not be of very great significance in practice. It is at least plausible that other potential sources of difficulty in the regulatory process dwarf the consequences of the distortion in the capital-labor ratio that the model predicts.

1. The A-J model of the regulatory process

■ The original Averch-Johnson analysis provides a simple and comprehensible depiction of a regulatory process and its relation to the behavior of the firm. The company is taken to produce one output and to use two inputs, labor and capital, each of which is available to it in unlimited quantities at a fixed price per unit. The firm's decisions are affected by the regulator in only one way: it is permitted to earn no more than some fixed proportion of the value of its capital—the regulatory "fair rate of return" on its rate base. In all other respects management is permitted to pursue its objective, maximization of profit, exactly as it could in the absence of regulation. Finally, it is assumed that the rate of return permitted to the firm by regulation is less than the return it would obtain if it were able to maximize its profits, but is at least as great as the rate of return on capital. These assumptions can be brought out more explicitly with the aid of mathematical notation, which largely follows that in the original A-J article.

---

[1] While the Wellisz model deals with the same sort of regulated firm, it concerns itself primarily with the effects of regulation on peak-load pricing rather than on input usage. The Wellisz discussion will, therefore, not be examined in this article.

Let

π = the firm's total profit,
z = the physical quantity of the firm's output,
x_1 = the quantity of capital in the firm's rate base,[2]
x_2 = the quantity of labor input,
r_1 = the unit cost of capital = the opportunity cost of resources tied up in plant and equipment,
r_2 = the unit cost of labor,
$s = r_1 + v$ be the rate of return permitted by regulation, $v \geq 0$,
$z = z(x_1, x_2)$ = the production function, $\partial z/\partial x_1 > 0$, $\partial z/\partial x_2 > 0$ and $z(0, x_2) = z(x_1, 0) = 0$,
$p = p(z)$ = the inverse demand function.

Then the company is assumed to select the values of z, x_1 and x_2 that maximize total profit

$$\pi = pz - r_1x_1 - r_2x_2 \tag{1}$$

subject to the regulatory constraint

$$\frac{pz - r_2x_2}{x_1} \leq r_1 + v = s\,. \tag{2}$$

Specifically, since $r_1 + v$ is taken to be less than the profit-maximizing rate of return, it is assumed by A & J that the constraint (2) is an equation rather than an inequality.[3]

Notice the very rudimentary regulatory process that is encompassed in this model. First, it is assumed that the firm can decide for itself on any price-output combination it wishes, so long as the rate-of-return constraint is satisfied. Second, it will be assumed at least for part of the discussion (following most of the literature) that the rate-of-return constraint is satisfied as an equality at all times and that there is no regulatory lag. If the regulatory commission decrees that net returns will not exceed s percent of the rate base, then they will equal s percent precisely, no more and no less, even as a temporary matter. Clearly, this is not quite the way the regulatory process works, though the model's simplification of the facts is not unreasonable as a first approximation. We shall see later what analytical difficulties one gets into in seeking to represent the regulatory process more closely.

The firm in this model is also rather rudimentary, its production process involving only two inputs and one output, and its objective being total profit maximization pure and simple. However, the choice of objective aside, no one seems to have suggested that any significant loss of generality results. The analysis also adopts several common assumptions as a matter of mathematical convenience: all relevant functions are assumed twice differentiable, and the appropriate second-order maximum (concavity-convexity) conditions are taken to hold throughout. Some of these will be discussed explicitly later.

---

[2] Because of depreciation the value of such resources will normally be less than the acquisition cost of these capital assets. For simplicity we will follow Averch and Johnson in assuming depreciation to be zero and the acquisition cost of capital to be unity.

[3] If capital were to depreciate, the rate of return in equation (2) would have to be expressed, not in terms of the acquisition cost of the capital, but in terms of its depreciated value. That is, x_1 would have to be the depreciated value of the company's capital, and current depreciation would appear as an operating cost. As just stated, it is convenient to assume this complication away.

2. The A-J theorems

■ Under the circumstances postulated it is easy to prove that the capital-labor ratio chosen by the firm will, in general, be different from the one that minimizes the cost of whatever output is decided upon. It is only slightly more difficult to show that the resulting input proportions will be more capital-intensive than efficiency requires. There has been some discussion of the validity of the argument used by A & J in arriving at this second result. As a matter of fact, as we shall see presently, it may not even be the result they had in mind when they wrote their paper. For such reasons, we find it useful to divide the hypotheses explicit and implicit in the discussion into several different propositions. We begin with

PROPOSITION 1. *The firm described in the model (1) and (2) will adopt input proportions different from those that minimize the cost of the final output level,* z^*.

PROOF: Minimization of the cost of the output selected requires minimization of $r_1x_1 + r_2x_2$ subject to $z(x_1, x_2) = z^*$, which yields the Lagrangian

$$r_1x_1 + r_2x_2 + \mu[z^* - z(x_1, x_2)]$$

and this leads to the first-order conditions

$$r_1 - \mu z_1 = 0\,, \quad r_2 - \mu z_2 = 0$$

where z_i denotes $\partial z/\partial x_i$, the marginal product of $x_i (i = 1, 2)$ or

$$r_1/r_2 = z_1/z_2\,. \tag{3}$$

Turning next to the regulated firm, its Lagrangian is, by (1) and (2) (treating the latter as an equality until otherwise noted),

$$L = pz - r_1x_1 - r_2x_2 - \lambda(pz - sx_1 - r_2x_2)\,. \tag{4}$$

We now obtain the first-order conditions[4]

$$(1 - \lambda)MR_z z_1 - r_1 = -\lambda s, \text{ i.e., } (1 - \lambda)MR_z z_1 - (1 - \lambda)r_1 = \lambda(r_1 - s) \tag{5a}$$

$$(1 - \lambda)MR_z z_2 - (1 - \lambda)r_2 = 0 \tag{5b}$$

where

$$MR_z = \frac{dpz}{dz} \text{(marginal revenue of output)}\,.$$

We will follow the literature in assuming through most of the discussion that

$$s > r_1 \tag{6}$$

(the "fair rate of return" exceeds the cost of capital). It then follows from (5a) that $\lambda \neq 1$. For if $\lambda = 1$ (5a) becomes simply $r_1 = s$, thereby contradicting (6). Thus we may divide equation (5a) by (5b), obtaining, after some rearrangement,

---

[4] Averch and Johnson utilize the Kuhn-Tucker conditions at this point because the regulatory constraint (2) is an inequality. However, they always assume this constraint is effective, so that they treat it as an equation throughout their discussion. As Takayama [10] points out, it is consequently unnecessary to use the Kuhn-Tucker conditions.

$$z_1/z_2 = r_1/r_2 - \lambda(s - r_1)/(1 - \lambda)r_2 . \tag{7}$$

Thus, if we can show that the last term in (7) is not equal to zero, it will follow that requirement (3) for the least-cost input combination is violated by our regulated firm. Because we have $s > r_1$ and $\lambda \neq 1$ as was just shown, we need only to show that $\lambda \neq 0$. This follows from comparison of the Lagrangian (4) with the original objective function (1). For if $\lambda = 0$ the two expressions are identical, so that the constrained maximization problem will yield a profit level identical with that which can be achieved by a firm unconstrained by regulation. Since we have assumed that the rate-of-return constraint does reduce company profits, it follows that we cannot have $\lambda = 0$. This completes the proof of the first A-J theorem.

A number of the published discussions leap from this result to one of the following two conclusions:

(ALLEGED) PROPOSITION 2. *The capital-labor ratio of the regulated firm will be larger than that of the unconstrained profit-maximizing monopolist.*

PROPOSITION 3. *The capital-labor ratio of the regulated firm will be larger than the one that minimizes costs for the output level that it elects to produce.*

It is not quite clear which of these A & J had in mind when they wrote their article. Certainly their discussion on this point is ambiguous (compare their Figure 1 with their discussion on p. 1053, top, and 1056, bottom).

As we shall see, proposition 2 is simply false, as a generalization, while proposition 3 is correct under the circumstances postulated although details of the proof require some care. We shall, in section 6, provide a counterexample to proposition 2 and discuss its problems in general terms. For the moment we shall confine our discussion to proposition 3, which is the more significant of the two from the point of view of efficiency of resource utilization. We will therefore present in detail a proof of the result. A critical step in the argument is the following lemma, which we will also utilize later in this article.

LEMMA. *If the firm maximizes its total profit (1) subject to the regulatory (inequality) constraint (2), in which* $s > r_1$, *and if, in addition, the regulatory constraint is binding and* $x_1 > 0$ *and* $x_2 > 0$, *then we must have* $0 < \lambda < 1$.

Since the discussion of this lemma in the literature is somewhat disjointed and impressionistic (see Takayama's criticism of A & J on this score), a more systematic derivation seems appropriate. Maximization of (1) subject to (2) yields as its first-order (Kuhn-Tucker) conditions with $x_1 > 0$, $x_2 > 0$, equations (5a), (5b), and

$$pz - sx_1 - r_2x_2 \leq 0 \tag{5c}$$

$$\lambda(pz - sx_1 - r_2x_2) = 0 \tag{5d}$$

$$\lambda \geq 0 . \tag{5e}$$

Writing R_1 as an abbreviation for $MR_z z_1$, the marginal revenue product of x_1, and using similar notation for x_2, we obtain by solving[5] (5a) for λ,

$$\lambda = (r_1 - R_1)/(s - R_1) . \tag{8}$$

The unconstrained profit maximum is given by the values of x_1, x_2 satisfying

$$R_1 = r_1 , \quad R_2 = r_2 .$$

By (5b), the second of these conditions is satisfied also by the regulated firm. Hence, since the regulatory constraint is assumed to be effective, i.e., it prevents the firm from maximizing profits, we must have

$$R_1 \neq r_1$$

and consequently, by (8), $\lambda \neq 0$. Therefore by (5e)

$$\lambda > 0 .$$

We have already seen, in the discussion immediately following (6), that $\lambda \neq 1$. To prove that $\lambda < 1$ we utilize the second-order conditions for the constrained profit maximum, which require a positive value for the bordered Hessian, i.e.,

$$H = \begin{vmatrix} (1-\lambda)R_{11} & (1-\lambda)R_{12} & s - R_1 \\ (1-\lambda)R_{21} & (1-\lambda)R_{22} & r_2 - R_2 \\ s - R_1 & r_2 - R_2 & 0 \end{vmatrix} > 0$$

where R_{ij} denotes $\partial^2 pz/\partial x_i \partial x_j$. However, by (5b) $r_2 - R_2 = 0$. Therefore the preceding condition becomes (on expanding the determinant in terms of its last column),

$$H = -(s - R_1)^2 (1 - \lambda) R_{22} > 0 .$$

Diminishing revenue returns to labor [concavity of $R(x_1, x_2)$] gives us $R_{22} < 0$ so that we obtain immediately

$$(1 - \lambda) > 0$$

as was to be shown.

It follows, incidentally, that $R_1 < r_1 < s$, i.e., that for the regulated firm in equilibrium the marginal revenue product of capital will be less than r_1, the cost of capital. This follows from (8), which indicates that if $r_1 < R_1 < s$ then $\lambda < 0$, while if $r_1 < s < R_1$ then $\lambda > 1$.

We now return to the main algebraic argument adduced in support of proposition 3. We see at once from the preceding lemma that the last term in (7) must be negative since $0 < \lambda < 1$ and since, by (6), $s > r_1$. From this one obtains the required result. Having shown by (7) that

$$z_1/z_2 < r_1/r_2 , \tag{9}$$

comparison with (3) demonstrates that the relative use of capital must, under (9), exceed the ideal proportion corresponding to (3). For z_1/z_2, the marginal rate of substitution of capital for labor, is now below the ratio of input prices r_1/r_2. With diminishing marginal rate of substitution of capital for labor for a fixed output level, this can occur only as the result of a relative increase in the use of capital.

---

[5] The division by $s - R_1$ is legitimate because if $s = R_1 = MR_z z_1$, we must have, by (5a), $s = r_1$, and this is ruled out by (6).

We deal next with the remaining contention of the A-J paper:

(ALLEGED) PROPOSITION 4. *The regulated firm will produce an output larger than that which maximizes profits.*

Strangely, one looks in vain for any attempt at a formal proof of this assertion. A & J seem to arrive at their result through a simple heuristic argument: From (5a) or (9) they conclude that the regulated firm will purchase more capital than the amount that maximizes profit [actually this does not follow directly from the equations cited, though, as will be shown later (proposition 5), the assertion about the capital stock *is* correct]. Similarly, equation (5b) can be rewritten simply as $MR_z z_2 = r_2$, asserting that the marginal revenue product of labor (x_2) will be equal to the wage level (r_2). This seems to have suggested that the profit-maximizing quantity of labor will be hired. With that quantity of labor and with the quantity of capital greater than the amount that maximizes profit, it follows that output must be greater than the profit-maximizing value of z.

As already noted, it is not clear that this is the argument that leads A & J to their allegation about the firm's output level. In any event, the argument is simply incorrect. Condition (5b) does *not* require the quantity of labor hired by the regulated firm to be the same as the amount that would be employed under pure profit maximization. The quantity of labor employed by the firm will still depend on the quantity of capital it uses and the nature of the production function. As Takayama [10] shows, rate-of-return regulation can in fact lead to a decrease in the quantity of labor hired by the firm, a point to which we will return presently. Indeed we will show later that, as a generalization, proposition 4 simply is not valid.

The invalidity of proposition 4 may perhaps be considered unfortunate. If it were true it might have been claimed as a virtue of the regulatory process. For generally the monopolistic firm's output is taken to be smaller than the level that is socially optimal. If regulation of rate of return were to guarantee an increase in output above the monopolistic level it might have been considered to be most desirable. Indeed, some authors have defended fair-rate-of-return regulation on the ground that it will have a beneficial effect on output. For example, Alfred E. Kahn [6] recognizes explicitly the input inefficiency such regulation can engender. But, he argues, by inducing the regulated firm to overinvest in capacity and hence to *increase its output,* fair-rate-of-return regulation helps to reduce—if not overcome—the monopolist's tendency to underinvest and restrict output. The fact that proposition 4 is not valid as a general statement weakens somewhat the case of Kahn and others who take a similar position.

3. The geometry of the A-J propositions

■ As it turns out, a good deal of illumination can be contributed by translation of the model's relationships into geometric terms. Two closely related graphic interpretations have been provided independently by Zajac [14] and by Klevorick and Baumol [8]. The discussion here will employ elements of both analyses.

FIGURE 1

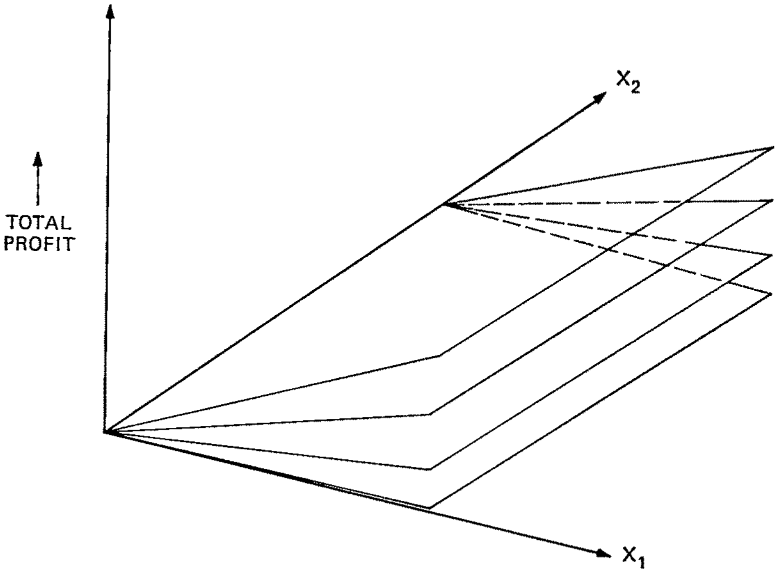


Figure 1 is a three-dimensional diagram depicting on its axes the values of x_1, x_2, and total profit. It also shows several planes, all of them hinged on the x_2 axis and hence satisfying the equation

$$\pi = vx_1 \tag{10}$$

for different values of v. Any such plane represents the rate-of-return constraint (2) for a particular value of v. The interpretation of the left-hand side of (10) as a profit function follows directly from (1) and (2). Rewriting our regulatory constraint (2) as an equality, as we are now treating it, and substituting this into the original profit function (1) we obtain at once

$$\pi = (pz - r_2x_2) - r_1x_1 = sx_1 - r_1x_1 = vx_1\,, \tag{11}$$

FIGURE 2

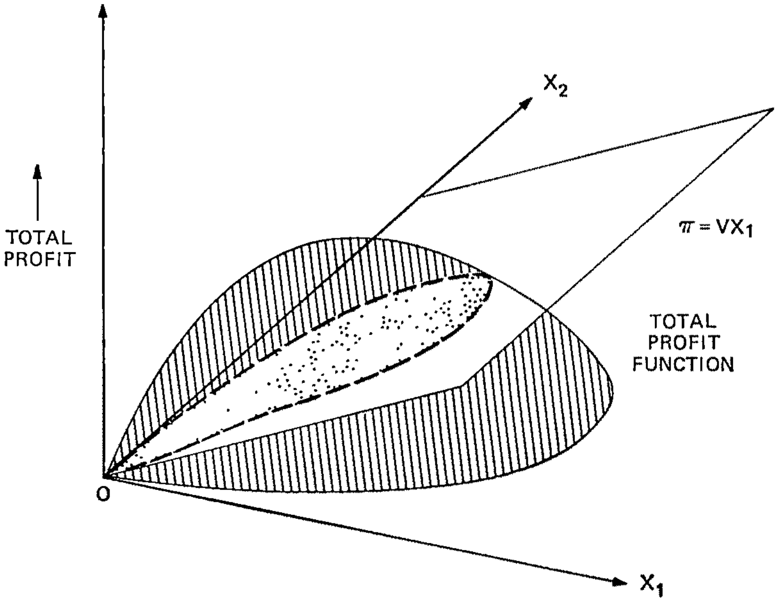
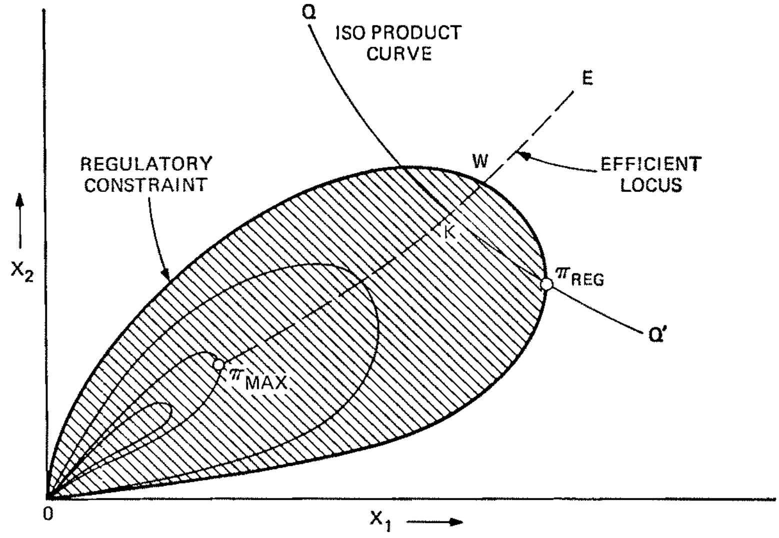


FIGURE 3



which is an expanded restatement of (10). We see from (11) that the slope of any of the cross sections of a constraint plane taken parallel to the (x_1, π) plane represents $s - r_1$, the excess of the regulatory "fair rate of return" over the cost of capital.

In Figure 2 the firm's profit function has been included in the diagram as the shaded hill-shaped figure. The heavy broken curve is the intersection of the profit surface and the regulatory constraint plane. Its projection on the x_1x_2 plane represents all combinations of the two inputs which just satisfy the rate-of-return constraint requirement. The heavy curve in Figure 3 depicts that projection on the x_1x_2 plane. The shaded region inside the constraint curve (the maximal-rate-of-return curve) corresponds to input combinations that yield rates of return higher than the maximum permitted by regulation. The heavy constraint curve is an iso-rate-of-return locus; i.e., points on that curve represent input combinations all offering the same rate of return on capital. Three other iso-rate-of-return loci, all of them offering higher rates of return than permitted by the constraint, are also shown. Point $\pi_{\max}$ corresponds to the highest point on the profit surface: it is the point the profit-maximizing firm would select in the absence of regulation.[6] The point π_{reg} is the input combination that maximizes profit given the regulatory constraint.[7] A moment's thought shows that π_{reg} must be the rightmost point on the regulatory constraint curve since this gets the firm to the highest total profit point on the constraint plane in Figure 2. Looked at another way, the rightmost point maximizes *total* profit consistent with the constraint because all points on the constraint curve yield the same net *rate* of return, v, and the rightmost point involves the largest rate base, x_1, and hence the largest value of $\pi = vx_1$.

---

[6] Note that point $\pi_{\max}$ which yields maximum total profits does not offer the maximal rate of return For, as the diagram shows, there are iso-rate-of-return loci inside the curve through $\pi_{\max}$, whose rates of return are higher than that through $\pi_{\max}$.

[7] While in this diagram the output at π_{reg} is greater than that at $\pi_{\max}$ as proposition 4 claims is generally true, we shall show presently that this result does not always hold.

The broken curve in Figure 3 is the locus of efficient combinations of labor and capital. For any given output level the least-cost input combination will be given by some point on this curve, $\pi_{\max}E$.

Proposition 3 corresponds to the assertion that the constrained maximum point lies below (to the right of) the efficient locus, $\pi_{\max}E$. For if this is true then QQ', the iso-product curve through point π_{reg}, will cross the efficiency locus at some point K that lies above and to the left of π_{reg} (since iso-product curves must have a negative slope in capital-labor space). K is, by definition of the efficient locus, the least expensive input combination for the production of the output level corresponding to QQ'. That is, K produces the same output as π_{reg} but at lower cost, and K clearly involves more labor and less capital than π_{reg}. Thus, if we can prove that π_{reg} lies below and to the right of the efficiency locus, it follows that the regulated firm in our model will utilize an inefficiently high capital-labor ratio to produce the output level it selects, as proposition 3 requires.

It therefore only remains for us to show that point π_{reg} does not lie on or above the efficiency locus $\pi_{\max}E$. To prove this, assume the contrary, as in the case depicted in Figure 4, where π_{reg} lies above the efficient locus. Let point A be the intersection point of the efficient curve with QQ', the iso-product locus through π_{reg}, and let B be the intersection of the efficient locus with the constraint curve.

Obviously, A is more profitable than π_{reg} since it yields the same output at a lower production cost. Moreover, by the concavity conditions implicit in the argument, profit increases monotonically as one moves along the efficiency locus toward $\pi_{\max}$. Thus B must be at least as profitable as A. Hence B must be more profitable than π_{reg}, contradicting the definition of π_{reg} as the most profitable point on the constraint locus.

This completes the Zajac proof of proposition 3.

There is an alternative geometric derivation of the A-J result that is also illuminating. Figure 5a depicts the firm's iso-profit curves—curves of constant *total profit*, the contour lines for the total profit surface that was represented in Figure 2. The concavity of the profit

FIGURE 4

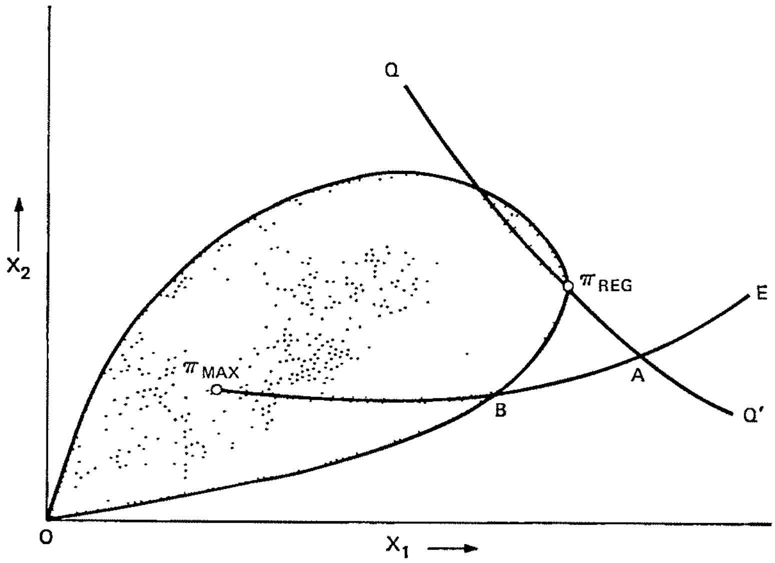
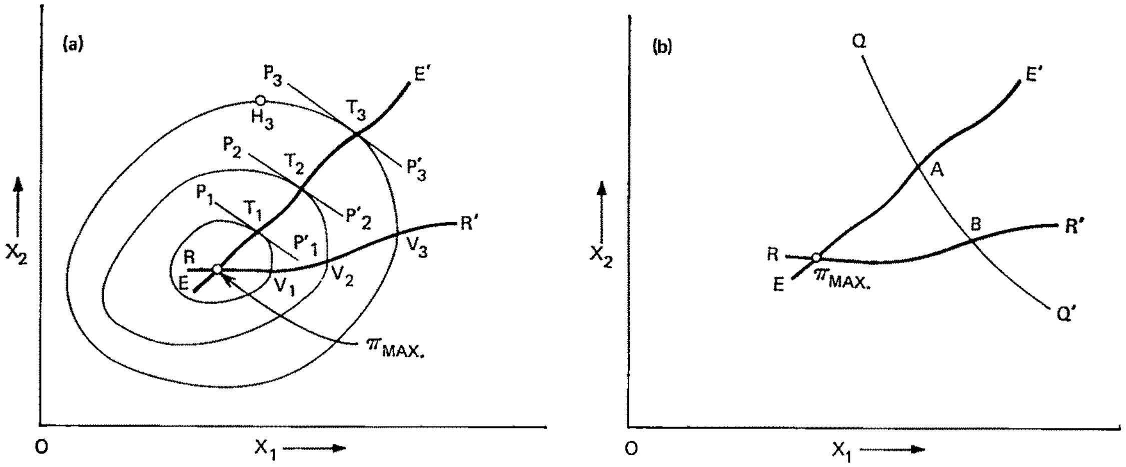


FIGURE 5



surface assures us that each such curve is the boundary of a convex region.

We shall use this diagram to describe explicitly the location of the firm's efficient locus, *EE'*, and of its expansion path *RR'*, showing the combinations of inputs that will be hired by the regulated firm with different values of *s*, the permitted rate of return. That is, *RR'* is the locus of all equilibrium points, π_{reg}, for our firm. If we can show that everywhere to the right[8] of the maximum profit point, $\pi_{\max}$, the locus *RR'* lies below and to the right of the efficient locus, as it does in Figures 5a and 5b, then the A-J result must follow. That is, under rate-of-return regulation, in producing a given quantity of output the firm must utilize a capital-labor ratio larger than that required for maximal efficiency. For since (Figure 5b) the iso-product curve *QQ'* has a negative slope, its intersection point, *B*, with the lower expansion path locus, *RR*, must lie below and to the right of *A*, the efficient point for that output.

We prove our assertion about the relative positions of *RR'* and *EE'* as follows: Curve *EE'* is the locus of tangency points,[9] T_i, between the iso-profit curves and the parallel negatively sloped price lines P_iP_i'. Hence curve *EE'* must intersect the iso-profit curves in their negatively sloping segments such as H_3V_3 in Figure 5a.

However, the expansion path, *RR'*, is the locus of right-hand vertical points on the iso-profit curves.[10] Our result follows at once from the convexity properties of these curves: A tangency point, T_i

---

[8] As is shown in proposition 5, below, when the rate of return constraint is effective, the firm will always use a quantity of capital greater than that at $\pi_{\max}$. Hence, any solution point must lie to the right of $\pi_{\max}$ in the diagram. This also follows directly from Figure 3 where we see that π_{reg}, the constrained maximum point, must always lie to the right of $\pi_{\max}$ since π_{reg} is the rightmost point on the constraint curve

[9] The slope of the price lines is, of course, equal in absolute value to the ratio of input prices, r_1/r_2 The slope of an iso-profit curve (see the following note) is $-\pi_1/\pi_2$ where $\pi_i = \partial\pi/\partial x_i$. Hence, tangency between an iso-profit curve and a price line requires

$$r_1/r_2 = \pi_1/\pi_2 = (MR_z z_1 - r_1)/(MR_z z_2 - r_2) \quad \text{or} \quad r_1 MR_z z_2 - r_1 r_2 = r_2 MR_z z_1 - r_1 r_2,$$

i.e., $r_1 z_2 = r_2 z_1$, which is the same as the necessary condition for efficiency (3).

[10] *Proof* The iso-profit curve is given by π = constant, from which we deduce $\pi_1 dx_1 + \pi_2 dx_2 = 0$, that is, the slope of the iso-profit curves is $dx_2/dx_1 = -\pi_1/\pi_2$. But from (5b) we obtain $\pi_2 = MR_z z_2 - r_2 = 0$. Hence, equilibrium of the regulated firm thus requires the iso-profit curve to be vertical.

An argument that is more geometric in spirit is also possible. We have already noted that equilibrium of the regulated firm requires it to be on the vertical (right-hand) point of an iso *rate-of-return* curve π/x_1 = constant, since at this point x_1, and hence π, will be at a maximum. The slope on the negatively sloping portion of any one of these curves on its rightmost end must lie above and to the left of the corresponding vertical (right-hand) point, V_t.

4. The cost of capital vs. the fair rate of return

■ A & J emphasize that their results depend on the assumption that the regulatory "fair rate of return" exceeds the cost of capital. This assertion is not very easy to test empirically, since the cost of capital presumably varies from firm to firm, depending at least in part on the degree of risk incurred by anyone who provides capital to it, and we do not really know how to measure that cost of capital with any high degree of precision. The proceedings of the regulatory commissions that determine the firm's fair rates of return certainly do have the appearance of a search for a figure approximating the cost of capital. They argue in terms of the return necessary to enable the firm to attract the capital it needs, and that is presumably what is meant by the (opportunity) cost of capital. In other words, it is at least arguable that the regulator often attempts to set the "fair rate of return" at a level as close to the cost of capital as he can determine.

For this reason, if not as a purely analytic matter, it is of interest to examine what happens to the A-J propositions as the regulated rate of return, s, approaches the cost of capital, r_1, and what occurs when the two coincide. As already implied, A & J themselves point out that their result breaks down when $s = r_1$. However the nature of this breakdown is a bit surprising. Our intuition might suggest that once capital no longer receives a return that may be termed "excessive," the firm will find it unprofitable to use an inefficiently large quantity of capital and it will therefore be motivated to utilize the efficient capital-labor ratio. As a matter of fact, when $s = r_1$ the firm has no motivation to select this ratio or any other particular ratio. *Any* capital-labor ratio becomes as profitable to the regulated firm as any other.

This result can be demonstrated with the aid of the first-order maximum conditions (2), (5a), and (5b). When $r_1 = s$ we see that $\lambda = 1$ satisfies (5a) and (5b). But when $\lambda = 1$ the firm's capital-labor ratio is indeterminate because the only first-order condition that continues to be effective is the original profit constraint (2). Any values of x_1 and x_2 that satisfy the regulatory constraint must, therefore, satisfy all of the first-order maximum conditions, and any such pair of values is as profitable to the regulated firm as any other.

Geometrically we can see more clearly why the A-J results collapse in this peculiar way. Returning to Figure 1, we recall from (10) and (11) that the slope of one of the constraint planes taken parallel to the x_1, π plane is $v = s - r_1$. Hence when $s = r_1$ the plane is flat—regulation offers the firm no opportunity to increase its total profit either through an increase or a decrease in the quantity of x_1. The constraint plane is now perfectly horizontal and, in fact, it coincides with the floor of the diagram—the x_1x_2 plane. Regulation

---

of the iso-rate-of-return curve is obtained by total differentiation thus:

$$d(\pi/x_1) = \left(\frac{\pi_1}{x_1} - \frac{\pi}{x_1^2} \right) dx_1 + \left(\frac{\pi_2}{x_1} \right) dx_2 = 0$$

so that

$$dx_2/dx_1 = -(\pi_1 - \pi/x_1)/\pi_2 .$$

At a vertical point on this curve we must therefore have $\pi_2 = 0$, which is precisely the condition for the iso *total* profit curve to be vertical since, as we have just seen, its slope is $dx_2/dx_1 = -\pi_1/\pi_2$.

now guarantees that the firm receives for its capital exactly what has to be paid for it, no more and no less. That is, profits will equal zero after payment of the "normal" return to capital, r_1. In this circumstance the firm has no particular motivation to employ an efficient quantity of capital, nor any motivation to avoid doing so. Entrepreneurial choice in this matter just has no effect on its total profits.

However, it is implausible that in such circumstances management will not adopt some other objective, one which is perhaps secondary but which, if profit is not affected, will be used as a basis for company decisions. It can then be shown that if $s = r_1$ and the firm seeks to maximize either total physical volume[11] or sales (total revenue), pz, it is likely to be driven to the ideal input combination point for the corresponding constraint locus, point W in Figure 3. That is, a regulatory restriction on rate of return with $s = r_1$, together with a company goal of sales or output maximization, will yield both ideal output and efficiency in use of resources!

The reason this is so is that maximization of pz or z does require the firm to be on the efficiency locus,[12] and the firm will therefore be forced to W, the only such point on our constraint curve.

Once we leave the case where r_1 and s coincide, matters seem to follow more closely the course one might expect. As A & J point out, if $s < r_1$ the firm loses money and it must, in the long run, go out of business. If, on the other hand, $s > r_1$, propositions 1 and 3, the basic results, follow directly. But even here things do not behave quite as one might at first anticipate.

It may seem plausible that the degree of inefficiency in the capital-labor ratio will decline as s is driven closer to r_1. That is, it may appear that the more closely the regulator succeeds in pushing the regulatory ceiling return, s, to the cost of capital, the less motivation the firm will have to use inefficiently large quantities of capital. In fact, this turns out to be false. On the contrary, one can show that, at least for some classes of production function, as s approaches r_1 the capital-labor ratio will depart further and further from its efficient value! More generally, one can show, following a conjecture of F. M. Scherer, that as s approaches r_1 the constrained profit-maximizing quantity of capital will *always increase.*

This result was proven geometrically by Fred M. Westfield,[13] but it was left to Takayama to provide the first rigorous mathematical derivation of the proposition. Curiously though, neither Westfield nor Takayama quite realized the importance of the proposition for regulatory policy. We shall return to its implications after the proof is completed.

Takayama's argument proceeds as follows. Let r_m be the rate of return on capital obtained at the (unconstrained) profit-maximizing

---

[11] The idea of maximization of physical volume was suggested to us by Milton Kafoglis.

[12] This can be shown as follows. Since the firm cannot survive with negative profits, the regulatory constraint *must* be in our present case satisfied as an equality, i.e., we must have $\pi = pz - r_1x_1 - r_2x_2 = 0$. Maximizing z subject to this constraint we obtain the Lagrangian conditions

$$z_1 + \lambda MR_z z_1 - \lambda r_1 = 0\,, \quad \text{i.e.,} \quad z_1(1 + \lambda MR_z) = \lambda r_1$$
$$z_2 + \lambda MR_z z_2 - \lambda r_2 = 0\,, \quad \text{i.e.,} \quad z_2(1 + \lambda MR_z) = \lambda r_2\,,$$

and straightforward division of one relationship by the other shows that the efficiency condition $z_1/z_2 = r_1/r_2$ must be satisfied. The proof for the sales maximization case where the maximand is pz rather than z follows the same argument step by step.

[13] In [12], Westfield's argument employed a geometric apparatus totally different from the one presented in section 3. For example, his diagram is drawn in capital-profits space rather than capital-labor space. See [12], pp. 429–32.

input-output combination. Then we have $r_m > s > r_1$. For this range of values of s we can conduct a comparative statics analysis of the value of dx_1/ds. We will prove

PROPOSITION 5. *For the firm that seeks to maximize total profit (1) subject to the regulatory constraint (2) we have* $dx_1/ds < 0$ *for* $r_1 < s < r_m$.

In other words, the greater the difference between the regulatory fair rate of return and the cost of capital (since we are increasing s holding r_1 constant), the *smaller* will be the value of x_1, the firm's use of capital!

This result is surely counterintuitive, at least at first blush.

The proof is not difficult. It follows from total differentiation of the first-order conditions (5a), (5b), and (2), utilizing the standard comparative statics procedure to determine how the values of x_1 and x_2 respond to a change in the value of s, holding r_1 constant. For simplicity, let us again rewrite total revenue pz as $R(x_1, x_2)$ so that, e.g., $R_1 = \partial R/\partial x_1$ represents the marginal revenue product of x_1. Then the first-order conditions (5a), (5b) and (2) become (treating the regulatory constraint as an equality)[14]:

$$R_1 - r_1 = (r_1 - s)\lambda/(1 - \lambda) \tag{12}$$

$$\begin{aligned} R_2 - r_2 &= 0 \\ R - r_2x_2 &= sx_1 . \end{aligned} \tag{13}$$

Differentiating the last of these conditions totally, one obtains

$$(R_1 - s)dx_1 + (R_2 - r_2)dx_2 = x_1ds . \tag{14}$$

By (13) the second term in this equation is zero. Moreover, by (12)

$$R_1 - s = r_1 - s + (r_1 - s)\lambda/(1 - \lambda) = (r_1 - s)/(1 - \lambda) . \tag{15}$$

Thus, substituting this relationship into (14) and after dropping its second term by (13) we have

$$dx_1/ds = (1 - \lambda)x_1/(r_1 - s) < 0 . \tag{16}$$

This is negative since the lemma of Section 2 gave us $0 < \lambda < 1$ and since $r_1 < s$ by assumption. Q.E.D.

It is worth examining this result somewhat more carefully. As already indicated, Takayama asserts that this is an alternative proof of the A-J theorem, but that is simply incorrect. Neither proposition 2 nor 3 follows from his argument. True, he has shown that the imposition of the constraint, as it reduces the rate of return from the profit-maximizing level r_m to the legal ceiling s, does increase the firm's use of x_1. But provided that the firm simultaneously expands its use of x_2 sufficiently, there is no need for the resulting capital-labor ratio to be inefficient. Indeed, as the value of s decreases we would normally expect an *efficient* value of x_1 satisfying the regulatory constraint to increase as well. For by increasing x_1 and x_2 sufficiently along the efficient path, profit can be driven down to the level required by a lower value of s. In terms of Figure 3, a lower rate-of-return ceiling moves the constraint curve further away from π_{max}, and therefore it moves the profit-maximizing point, π_{reg}, the vertical point on the constraint curve, further to the right. That is precisely why a decrease in the permitted rate of return necessarily increases x_1, the quantity of capital that will be used by the firm. However, the reduction in value of s also leads the efficient point, W, on that curve to move further to the right, meaning that a larger use of capital is also required to maintain efficiency. In sum, there is no necessity of inefficiency implicit in the result that the value of x_1 is increased by regulation.

---

[14] Recall that since $r_1 < s$ we know that $\lambda < 1$ and since $s < r_m$ we have $\lambda > 0$ (the Lemma of section 2).

This moderately painstaking discussion of the Takayama result was undertaken primarily because of the importance of the proposition for regulatory policy. For proposition 5 at least yields the disquieting result that a better job of setting the fair rate of return close to the cost of capital will not reduce the *absolute* capital bias and that it need not reduce the inefficiency in the input proportions shown by proposition 3. On the contrary, the result of better estimation of the cost of capital by the regulator may simply be to aggravate that inefficiency if the rise in the use of capital is not matched by an appropriate increase in the quantity of labor utilized.

A concrete example will demonstrate that this can in fact happen, i.e., that as s approaches r_1 not only will x_1 increase but there can also be an increase in the difference between the profit-maximizing capital-labor ratio x_1/x_2 and the efficient value of that ratio.

For this purpose let the total revenue function be

$$R(x_1, x_2) = a \ln x_1 + b \ln x_2$$

where $a > 0$, $b > 0$ so that $R(x_1, x_2)$ is strictly concave. Then as our marginal revenue products we have

$$R_1 = a/x_1 \quad R_2 = b/x_2 .$$

By (3) the efficient path is given by $R_1/R_2 = r_1/r_2$, i.e., by

$$r_1 b x_1 = r_2 a x_2 ,$$

which is clearly a straight line through the origin in (x_1, x_2) space. Thus, in this case the efficient path involves a constant capital-labor ratio.

On the other hand, we know by (16) that $dx_1/ds < 0$ while, as we will see presently, from (18), dx_2/ds must in this case equal zero since $R_{21} = 0$. Thus we see that as s moves closer and closer to r_1 the constrained profit-maximizing x_1/x_2 ratio must always increase, and hence it must constantly depart farther from the constant proportion between x_1 and x_2 that is required for efficiency.[15]

5. The demand for labor and the profit-maximizing output level[16]

■ One may be tempted to infer, carelessly, that since by (5b) $R_2 = r_2$, the regulated firm will employ the optimal quantity of labor. In fact, this conclusion is simply incorrect. A glance at Figure 3 readily confirms that the firm in the model will employ an inefficiently small quantity of labor for whatever level of output it chooses to produce. The reason, of course, is that the condition $R_2 = r_2$ is consistent with the efficiency requirement $R_1/R_2 = r_1/r_2$ if and only if $R_1 = r_1$, which will occur only at the unconstrained profit-maximizing point.

---

[15] A more "realistic" example is provided by the revenue function $R(x_1, x_2) = 2ax_1^{\frac{1}{2}} + 4bx_1^{\frac{1}{4}}x_2^{\frac{1}{4}} + 2cx_2^{\frac{1}{2}}$, $a > 0$, $b > 0$, $c > 0$. The argument is similar to that just given.

[16] After the paper was completed it came to our attention that Westfield had, in fact, presented results essentially the same as those derived in this section (See [12], p. 430) His verbal statements involve some slight imprecision, and there appears to be an error—probably typographical—in the footnote where he demonstrates the effect of regulation on the profit-maximizing output level. However, there is no doubt that he had arrived correctly at the relevant conclusions.

Consequently, one may wish to inquire further into the effect of the regulatory constraint upon the firm's use of labor. Does the firm's employment of labor necessarily increase (or necessarily decrease) when a regulatory revenue constraint is imposed? As is the case in its use of capital, is there necessarily a monotonic relationship between the quantity of labor hired and the gap between s and r_1? It can be shown that no such generalization is possible. Depending on the marginal rate of substitution between x_1 and x_2, the employment of labor may either increase or decrease as s gets closer to r_1. If labor and capital are complementary in the gross revenue function, then as the quantity of capital used by the firm increases its use of labor will also rise. But if capital is a substitute for labor in producing revenue, then x_1 and x_2 will move in opposite directions in the A-J model.

The proof follows the comparative-statics procedure utilized in deriving the Takayama result. However, the proof will be indicated only implicitly as part of our discussion of another issue—the effects of the regulatory constraint on the firm's output level. It will be recalled that the A-J allegation on this subject, designated proposition 4 above, asserts that effective rate-of-return regulation will lead the firm to increase its output level. Let us now see under what circumstances this is true.

Using the production function $z = z(x_1, x_2)$, one obviously has

$$\frac{dz}{ds} = z_1 \frac{dx_1}{ds} + z_2 \frac{dx_2}{ds} \tag{17}$$

where z_1 and z_2 are, respectively, the marginal products of capital and labor. To determine the sign of dz/ds it is necessary to evaluate dx_1/ds and dx_2/ds. Total differentiation of (13) yields

$$R_{21}dx_1 + R_{22}dx_2 = 0 \quad \text{or} \quad dx_2 = -(R_{21}/R_{22})dx_1 . \tag{18}$$

Since $R_{22} < 0$ by the strict concavity of R or simply by diminishing (marginal revenue) returns to labor, it follows immediately that the change in the employment of labor, dx_2, must have the same sign as the change in the use of capital, dx_1, if and only if capital and labor are complements in the production of revenue, $R_{21} > 0$. If they are substitutes in revenue production, their employment levels must change in opposite directions.

Substituting from (18) for dx_2 into (17), one obtains

$$\frac{dz}{ds} = \left(z_1 - z_2 \frac{R_{21}}{R_{22}}\right) \frac{dx_1}{ds} = \frac{z_1 R_{22} - z_2 R_{21}}{R_{22}} \frac{dx_1}{ds} . \tag{19}$$

We want to evaluate $z_1R_{22} - z_2R_{21}$ in (19). For this purpose we note that since R = total revenue = pz,

$$R_2 = (p + p'z)z_2$$
$$R_{22} = (p + p'z)z_{22} + (2p' + p''z)z_2^2$$
$$R_{21} = (p + p'z)z_{21} + (2p' + p''z)z_1z_2 .$$

Consequently,

$$z_1R_{22} - z_2R_{21} = (p + p'z)z_1z_{22} + (2p' + p''z)z_2{}^2z_1 \\ - (p + p'z)z_2z_{21} - (2p' + p''z)z_2{}^2z_1$$

or

$$z_1R_{22} - z_2R_{21} = (p + p'z)(z_1z_{22} - z_2z_{21})\,. \qquad (20)$$

Now, $(p + p'z)$ is simply MR_z, the marginal revenue of z, and is usually assumed to be positive (though with a constraint on rate of return this, clearly, need not be so). We may also assume $R_{22} < 0$ (diminishing marginal revenue returns to labor—which follows from the strict concavity of R). As a result, by (20) and (19) dz/ds must have the same sign as dx_1/ds (i.e., they must both be negative) if and only if

$$z_1z_{22} - z_2z_{21} < 0\,. \qquad (21)$$

There is, however, no basis on which to prejudge the sign of z_{21} since capital and labor can be either complements or substitutes in producing output. Therefore, we simply cannot conclude that the A-J proposition 4 is always valid.

Yet there is a plausibility argument in its favor. Since we do expect $z_1 > 0$ and $z_2 > 0$ (positive marginal productivity of each input) we can write (21) as

$$\frac{z_{22}}{z_2} < \frac{z_{21}}{z_1}\,.$$

But this requires that the proportional change in the marginal physical product of labor which results from an increase in labor *alone* is *algebraically smaller* than the proportional change in the marginal physical product of capital which results from the same increase in labor alone. The condition is met if $z_{21} > 0$ so that capital and labor are complements in production. If, on the other hand, $z_{21} < 0$ so that capital and labor are substitutes, then it is met if the proportional decrease in marginal physical products as a result of a labor increase is greater for labor than for capital. In sum, the expression $z_1z_{22} - z_2z_{21}$ can be said to be negative so long as labor is a better substitute for labor than it is for capital.[17]

Alternatively, by differentiating totally the equilibrium conditions for an unconstrained profit-maximizer, that is, (5a) and (5b) with $\lambda = 0$, and the production function for the firm, all with respect to a change in the market cost of capital, r_1, one can easily show that the key expression in (19), $z_1R_{22} - z_2R_{21}$, will be positive if and only if x_1 (capital) is a "regressive" input—i.e., an input such that a *rise* in its price leads the firm, *ceteris paribus*, to *increase* the output of the final product. In short, if marginal revenue is positive and if labor is a better substitute for labor than it is for capital—or equivalently, if capital is *not* a regressive factor—then $dz/ds < 0$ so that output will indeed increase as the regulatory constraint becomes tighter (s decreases) as is required for the validity of the A-J conjecture (proposition 4 above).

---

[17] If both terms in (21) are multiplied by x_2, the expression takes the form of a comparison of elasticities of marginal products with respect to changes in x_2 and it is then units free.

6. The capital-labor ratio with and without regulation

■ Finally, we are now in a position to deal with (alleged) proposition 2, which claims that the profit-maximizing firm subject to an effective rate-of-return constraint will necessarily employ a capital-labor ratio higher than that of an unconstrained enterprise.

We will first show in general terms why this need not be true and then we will offer a specific counterexample.

Proposition 2 asserts that

$$\frac{d(x_1/x_2)}{ds} < 0 ,$$

that is, by (16)

$$\operatorname{sign} \frac{d(x_1/x_2)}{ds} = \operatorname{sign} \frac{dx_1}{ds} . \tag{22}$$

Since we care only about signs we can look at the *proportional* change in x_1/x_2 when s decreases (i.e., when regulation is tightened). We have as this proportional change

$$\frac{d(x_1/x_2)}{x_1/x_2} = \left(\frac{dx_1}{x_2} - \frac{x_1}{x_2^2} dx_2 \right) \frac{x_2}{x_1} = \frac{dx_1}{x_1} - \frac{dx_2}{x_2} .$$

But from (18) we know that

$$dx_2 = -\,(R_{12}/R_{22}) dx_1 .$$

Therefore

$$\frac{d(x_1/x_2)}{x_1/x_2} = \frac{dx_1}{x_1} + \frac{R_{12}}{R_{22}} \frac{dx_1}{x_2} = dx_1 \left(\frac{1}{x_1} + \frac{1}{x_2} \frac{R_{12}}{R_{22}} \right) .$$

Thus (22) holds if and only if

$$\left(\frac{1}{x_1} + \frac{1}{x_2} \frac{R_{12}}{R_{22}} \right) > 0 , \quad \text{i.e.,} \quad \frac{x_2}{x_1} > -\frac{R_{12}}{R_{22}} .$$

Now this is surely true if $R_{12} < 0$, i.e., if capital and labor are substitutes in the production of revenue, because then we have $R_{22} < 0$, $R_{12} < 0$.

But suppose $R_{12} > 0$. Then the preceding condition becomes

$$\frac{x_2}{x_1} > \frac{R_{12}}{R_{22}} = \frac{|R_{12}|}{|R_{22}|} \quad \text{or} \quad \frac{|R_{22}|}{x_1} > \frac{R_{12}}{x_2}$$

which is obviously not necessarily true, nor is it even easily interpretable.

We now provide a simple counterexample to the allegation of proposition 2. Let us suppose the firm's revenue function is

$$R(x_1, x_2) = a x_1^2/2 + b x_1 x_2 + c x_2^2/2$$

where

$$a < 0 , \quad c < 0 \quad \text{and} \quad ac - b^2 > 0$$

so that $R(x_1, x_2)$ is strictly concave. By (5b) our constrained profit maximizer will satisfy $R_2 = r_2$, i.e.,

$$b x_1 + c x_2 = r_2 \quad \text{or} \quad x_2/x_1 = r_2/c x_1 - b/c$$

which increases as x_1 increases since $r_2 > 0$, $c < 0$. But by proposition 5 x_1 will be larger for the constrained than for the unconstrained firm, so that in this case the labor-capital ratio will be *increased* by the regulatory constraint, contrary to the assertion of proposition 2.

7. The revenue-maximizing firm under regulation

■ We have now completed our discussion of the basic A-J propositions and their formal validity. We turn now to some related theoretical results that have developed from the model.[18] One line of investigation which has grown out of the A-J analysis has examined the effects of the regulatory constraint (2) upon the firm that seeks to maximize total revenue or some other maximand instead of total profits. Such models have been studied by Bailey and Malone [3], Kafoglis [5], and Zajac [14]. Among the maximands that have been suggested by these authors are total revenue (sales), pz, physical volume, z, and rate of return, π/x_1. In each case the analysis proceeds exactly as before, only with the substitution of one of the preceding maximands for the total profit function (1). There is no need to go through the details of the analyses.

Not surprisingly, Zajac concludes that for a regulated firm maximizing its rate of return, the capital-labor ratio becomes indeterminate. Any point on the constraint curve in Figure 3 becomes as good as any other. The firm will simply select *any* capital-labor combination that enables it to attain that iso-rate-of-return curve, since it will then have achieved the greatest rate of return permitted by regulation.

Bailey and Malone, however, following the A-J procedures find that under sales (i.e., total revenue) maximization the nature of the inefficiency discussed by Averch and Johnson appears to reverse itself. Assuming that the firm's production function is quasiconcave so that the iso-product curves are convex to the origin, they prove:

PROPOSITION 6. *The sales-maximizing firm under rate-of-return regulation is motivated to use a labor-capital ratio greater than that which minimizes cost for the output level it chooses to produce.*

To prove proposition 6 we note that to obtain the new maximand, total revenue, we simply drop the terms $-r_1x_1$ and $-r_2x_2$ from the profit function (1) so that (5a) and (5b) become respectively

$$(1-\lambda)MR_z z_1 = -\lambda r_1 - \lambda(s - r_1)$$
$$(1-\lambda)MR_z z_2 = -\lambda r_2 .$$

Dividing the first equation by the second we have

$$z_1/z_2 = r_1/r_2 + (s - r_1)/r_2 \tag{23}$$

where the last term is obviously positive. Comparison with (7) shows that, unlike the profit-maximizing case, this time z_1/z_2 exceeds r_1/r_2 which, given the convexity assumption for the iso-product curve, can only occur at a point where the labor-capital ratio exceeds its efficient value.

---

[18] We had intended to include in this paper a discussion of Klevorick's graduated fair return proposal [7]. However, C R. Wichers [13] has found a significant error in the original formulation and it has therefore been thought advisable to postpone the discussion until the matter has been resolved.

This is shown explicitly in Figure 3. Suppose the firm's output corresponds to iso-product locus QQ'. At every point on that locus, since output is fixed, the price of the company's product is the same. Hence at any point on QQ' the value of z_1/z_2 is given by the slope of this curve. At the efficient point, K, this slope is equal to r_1/r_2, the slope of the budget line. Only at points on QQ' above and to the left of efficient point K will z_1/z_2, the slope of QQ', exceed r_1/r_2. But at these higher points on QQ' the labor-capital ratio exceeds its efficient value.

To compare geometrically the sales-maximizing firm's behavior with that of the profit-maximizing firm, assume that demand is elastic over the relevant range. In this case the firm will move to the highest iso-product (iso-revenue)[19] locus attainable on the constraint locus.

Usually in sales (total revenue) maximization models the firm is taken to maximize its total revenue function $R(x_1, x_2)$ subject to some sort of profit constraint such as $\pi(x_1, x_2) \geq mx_1$, which guarantees that the firm take advantage of opportunities to earn enough profit to satisfy its stockholders. The discussion implicitly assumes that the profit constraint (2) of the regulatory model also serves, with the inequality reversed, as the *minimum* profit requirement for the sales-maximization hypothesis. If the regulatory constraint permitted the firm to earn more than the profit needed to satisfy stockholders, revenue maximization would require it to move to an iso-rate-of-return curve lying entirely outside the shaded region in Figure 3, where the shaded region represents the input combinations prohibited by the fair rate of return ceiling.

Since the firm wishes, therefore, to move to the highest iso-revenue locus on the constraint curve, it will go to the locus *tangent* to the constraint curve. And since any iso-revenue (iso-product) curve is negatively sloped, the input point chosen by the regulated sales maximizer will therefore occur on a negatively sloping portion of the constraint curve. It must then involve a labor-capital ratio higher than that selected by the profit-maximizing regulated firm which, as we have seen, corresponds to the rightmost (vertical) point on the constraint curve (point π_{reg} in Figure 3).

The preceding result, since its conclusion is precisely the opposite of that of proposition 3, may be considered one of a number of attempts to show that the A-J conclusions lack general validity.

A rather informal argument by Zajac with a somewhat similar objective may be mentioned in passing. This discussion raises no questions about the formal validity of the deductions from the theoretical model but does indicate doubts about the model's applicability. Zajac suggests that the firm in reality knows its cost relationships far better than its demand function. As a result it may find it far easier to locate points on its efficient locus than to approximate its constrained profit maximum point, π_{reg} in Figure 3. Zajac suggests that the firm may perhaps attempt to work towards π_{reg} by any of a large number of alternate trial-and-error procedures. He chooses to focus on two of these possible series of moves. In the first set the firm follows along the regulatory constraint curve, while in the other the firm, knowing no better, proceeds to the point of minimum cost for whatever output it happens to be producing initially (a point on the efficient path) and then moves roughly along the efficient path toward the most profitable point on that path, the point where it intersects the constraint curve. Which of these strategies, *if either*, the firm is in fact likely to adopt we cannot say. Zajac points out, though, that depending on the relative speed of the two trial-and-error processes, it is perfectly possible that the efficient-path strategy will actually prove more profitable for the firm, despite its relative unprofitability in static equilibrium with full information. Clearly, if this is the course selected by management, the conclusion of proposition 3 simply does not hold. Of course, the Zajac argument shows only that with imperfection of information of the sort that may plausibly be expected in this area it is virtually impossible to predict what will happen. Things may turn out either better or worse than the A-J model suggests. But that is just the point that Zajac wishes to make.

---

[19] Since we take $p = p(z)$, for any fixed value of z we must also have a constant value of p. Consequently, a locus along which z is fixed (an iso-product locus) must also involve a fixed total revenue (pz).

8. Regulatory lag

■ In practice the rate-of-return constraint is not enforced continually. Rather, the regulator restudies the company's prices and profits from time to time and decrees, if necessary, that these be readjusted so as to yield for the firm the fair rate of return. Regulatory lag means, for example, that the firm will often be able to earn profits greater than those considered appropriate by the regulator, as judged by the permitted rate of return, for limited periods of time. We will argue later that this lag is potentially of very great significance for economic efficiency in general. In the present context one might surmise that it serves to weaken the A-J input-proportion effect, at least to some extent.

Bailey and Coleman [2] have undertaken to investigate formally some of the implications of regulatory lag for the A-J model. They discuss two possible courses of events, one in which the firm simply acquires inefficiently large quantities of capital and holds them idle until the regulator, by changing the price the firm can charge, readjusts the total return the firm is permitted to earn, and the second in which the company makes the best possible use of that capital even though the quantity is inefficiently large for the output being produced. Since the former case, though the less realistic, is considerably easier to describe, and since it enables us to characterize the essence of the Bailey-Coleman argument, we will use it as the basis of our discussion. Moreover, although Bailey and Coleman considered the case of a firm with a finite horizon as well as the case of a firm with an infinite horizon, we shall restrict our discussion to the latter as it serves to bring out the basic point of their argument.

The course of events envisioned in the Bailey-Coleman article is roughly the following: capital can be acquired easily, but the regulator does not adjust the company's prices immediately after the acquisition of new capital. The firm must therefore consider the advisability of obtaining an inefficiently large quantity of capital à la A-J, knowing that the new capital will reduce profits until the price adjustment and increase profits thereafter. While management does not know the length of the time it takes for the regulator to react by making a price adjustment, it does have an estimate of its expected value, believing this to be n periods in the future.

Utilizing the same notation as before, if the excess capital is acquired at once, the company will immediately incur a cost of $r_1\Delta x_1$ dollars per period, where Δx_1 is the quantity of capital in question. Once the regulatory readjustment takes place the company will add to its total earnings $s\Delta x_1$ dollars per period but will continue to have to pay $r_1\Delta x_1$ dollars per period for the privilege.

Letting D be the appropriate discount factor, the net effect upon the profits of the firm will therefore be

$$\Delta\Pi = -r_1\Delta x_1 \sum_{t=0}^{n-1} D^t + (s - r_1)\Delta x_1 \sum_{t=n}^{\infty} D^t .$$

This will be positive, (i.e., the investment will be profitable) if and only if

$$-r_1 \sum_{t=0}^{n-1} D^t + (s - r_1) \sum_{t=n}^{\infty} D^t > 0 , \quad \text{i.e.,}$$

$$-r_1(1 - D^n)/(1 - D) + (s - r_1)D^n/(1 - D) > 0 ,$$

or

$$-r_1 + sD^n > 0 .$$

Since $D < 1$, the last inequality shows that if the length of the lag is sufficiently great relative to $(s - r_1)$ it will not pay to overinvest by the postulated amount Δx_1. Increased regulatory lag will therefore decrease the likelihood in this model that the firm will want to acquire an excessive quantity of capital in the manner suggested by A & J.

In this simple and unrealistic case, the one in which the excess capital is held idle, the amount of excess capital is not well defined because there are no diminishing returns to excess capital. The limit to the amount that will be acquired is presumably imposed by an influence that does not appear in the model explicitly. If the elasticity of demand for its product limits the amount that the company can, in fact, earn to less than the amount permitted by an expanded rate base, there simply will be no point in acquiring additional capital. The value of Δx_1 that enables the firm to reach the point of maximum profits under the fair rate of return constraint (π_{reg} in Figure 3) will then be the amount, Δx_1^*, that it pays the company to obtain. If n is sufficiently small to make it profitable to acquire any excess capital at all, it will pay to obtain this full amount, Δx_1^*.

Bailey and Coleman show, however, that this all-or-none result is an artifact of the simplifying premise that the excess capital is retained unused. They argue that it will, in fact, always pay to use the capital to replace some of the company's labor, rather than keeping it idle. In that case we may expect a diminishing profit yield to increased quantities of capital as a consequence of the diminishing marginal rate of substitution of capital for labor. Bailey and Coleman show that in their model it will always pay the firm to acquire *some* excess capital, but that the amount it is profitable to obtain will vary inversely with the length of the regulatory lag. The reason is simple. In their world the excess capital always faces a period in which it is expected to incur a net loss. The sooner it can begin to earn a net profit as a result of regulatory price readjustment, the more capital it will pay the firm to acquire for the purpose.

The authors of this paper have also constructed and examined a number of models of regulatory lag. The regulatory process and its effects as depicted in our model are, however, somewhat more complex than in most of the other models that have been described, in accord with our view of the nature of the regulatory process in reality. In our model the time sequence of profits and losses is reversed from that in the Bailey-Coleman model. While Bailey and Coleman regard the period before a regulatory review as the time when the firm suffers a loss because it is carrying an excessive amount of capital, now the period between reviews is regarded as the time when the firm has the possibility of earning a profit rate exceeding that specified by the constraint. When the regulatory review occurs, this excess is eliminated by the regulators' adjustment of the prices the firm can charge. In contrast, in the Bailey-Coleman model it is the review point which enables the firm to gain at last the benefits of the excess capital it has purchased earlier.

As we see it, regulation proceeds, roughly, by taking the latest information available on company costs and demands, and resetting prices so that the firm's *current* rate of return—based on its current capital stock, technology, demand conditions, and so forth—is reduced to the "fair" rate level. The firm is required to leave those prices unchanged until another regulatory review takes place either at the behest of the company or some other interested party. Thus, unlike the company in the A-J model, the firm in this analysis cannot simply select an output level and charge for it the corresponding price as indicated by the demand curve. On the contrary, the price is set by the regulators and the company, as a public utility, is then required to supply whatever volume of output happens to be demanded at that price.

How then can any excess profits be earned? Leaving aside any demand shifts the firm may be able to induce with the aid of advertising and the like, the answer is that by research and innovation the company can hope to reduce its costs below those that prevailed when prices were last set by the regulators. So long as the initial output and input prices prevail, it will be able to retain for its stockholders whatever savings the innovation makes possible. But this additional earnings flow is only a temporary benefit. At the next regulatory review, prices will be readjusted once again to take into account the firm's improved technology and the process begins all over again.

In the analysis of this process the firm is taken to utilize (at least) three inputs: labor, capital, and knowledge. Labor is a non-durable input, and the amount employed is determined anew at each of the firm's decision points. Each instant's net investment, however, constitutes an addition to the firm's total stock of capital, and, similarly, each period's research adds to the firm's accumulated stock of knowledge. The issues, then, are: first, whether or not in this model the regulated firm will be motivated to employ an efficient combination of the three inputs; second, whether the degree of input inefficiency, if there is any, varies with the length of the regulatory period; and third, how the equilibrium stock of knowledge is affected by the length of time between regulatory reviews.

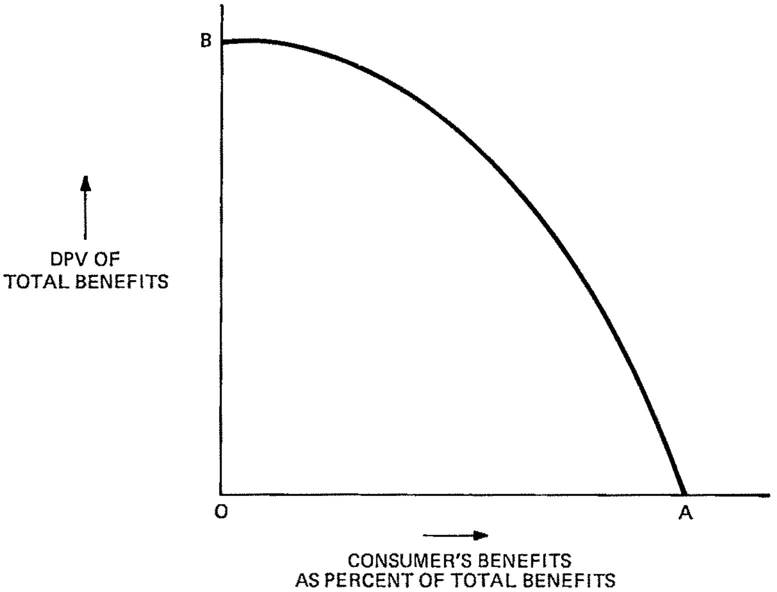
This last issue is potentially of great importance. If one can demonstrate a functional relationship between the length of time between regulatory reviews and the long-run equilibrium stock of knowledge, it would suggest that regulators have at their disposal an instrument, regulatory lag, of which they have not made much conscious use. Suppose, for example, that the regulated firm faces a static demand curve—i.e., one which does not shift with time. Now, the long-run goal of regulation is surely related to the net flow of benefits to consumers and to the producer (the total producer's plus consumers' surplus produced by the regulated firm) *and* the distribution of those benefits between producer and consumers. Other things being equal, the total net benefits generated by the regulated firm will be increased by downward shifts in the firm's long-run marginal cost curve produced by improvements in its technology—i.e., by increases in the long-run equilibrium stock of knowledge.

Suppose the equilibrium stock of knowledge increases with the length of the regulatory lag, because the firm undertakes more research the longer are the periods during which it enjoys the temporary benefit of an excess earnings flow. Then the total discounted present value of the benefits generated by the regulated firm will rise when the length of the regulatory lag increases. For example, with instantaneous regulation—a lag of length zero—and technical progress which increased the productivity of capital and labor proportionately, one would expect the firm to undertake no research outlays since it could never reap any benefits from such an expenditure, while with no regulatory review—a lag of infinite length—one would expect the firm to accumulate a maximal stock of knowledge since it reaps the benefits of its research work forever.

While one might consequently expect the length of the lag to be positively related to the *total* discounted present value of producer's plus consumers' benefits, it is clear that a smaller *proportion* of these benefits will be distributed to the firm the shorter is the lag. That is, the shorter the periods during which the firm can earn "excess" profits and retain them for itself, the greater is the portion of the accumulated surplus which is taken from the firm and given to consumers.

To each length of regulatory lag, then, there corresponds a discounted present value of total gains and a proportion of these gains which accrues to the consuming public. The latter decreases as the lag increases, while one would expect the former to increase as the lag increases. In the simplest case, society faces an opportunity locus of the type shown in Figure 6, where total surplus (benefit) is measured on the vertical axis and the proportion of the total gains accruing to consumers is measured on the horizontal axis. Point A corresponds to instantaneous regulatory adjustment, while point B corresponds to an infinitely long period between rate reviews. Society—or the regulators—may then be assumed to choose among the points on this locus with the aid of a social welfare function defined on the two arguments—total net benefits, and their distribution—and to maximize the social welfare function subject to the opportunities-locus constraint. In the simplest case, the optimal point selected in this way will correspond to a unique length of the regulatory lag. In any event, the regulators can then choose a length for the period between reviews which will attain the social optimum with respect to the amount and distribution of producer's plus consumers' surplus. If regulatory agencies have chosen the time period between

FIGURE 6



reviews in some more or less *ad hoc* fashion, this model would suggest that they have neglected a potentially valuable instrument of control.

With the aid of some additional notation, it is easy to describe one of the formal models used to investigate this issue of regulatory lag and research. Let

$$\left.\begin{array}{l} z = \text{the flow of output} \\ x_1 = \text{the stock of capital} \\ x_2 = \text{the quantity of labor used} \\ x_3 = \text{the stock of knowledge} \\ \dot{x}_1 = \text{the flow of investment} \\ \dot{x}_3 = \text{the flow of research activity,} \end{array}\right\} \text{all at time } t,$$

p_i = the price the firm is allowed to charge in the ith regulatory period: p_1 is exogenously given while for $i \geqq 2$, p_i emerges as the adjusted price after the $(i-1)^{\text{st}}$ regulatory review,

$z = F(x_1, x_2, x_3)$ is the production function,

r_2 is the wage rate,

$g(\dot{x}_1)$ is the (total) cost of investment,

$h(\dot{x}_3)$ is the (total) cost of research and,

v is the discount rate.

Assume that the length of time between regulatory reviews, i.e., the length of the regulatory lag, is T.

The discounted present value of the firm's profits is then

$$\Pi = \sum_{i=1}^{\infty} \int_{(i-1)T}^{iT} [p_i F(x_1, x_2, x_3) - r_2 x_2 - g(\dot{x}_1) - h(\dot{x}_3)] e^{-vt} dt . \quad (24)$$

The firm's goal is to maximize the discounted present value of profits subject to the transition equations (25) and (26) for capital and the stock of knowledge:

$$\frac{dx_1}{dt} = \dot{x}_1 \quad (25)$$

$$\frac{dx_3}{dt} = \dot{x}_3 \geq 0 . \tag{26}$$

As is clear from (25) and (26), we assume that there is no capital depreciation and no decay of knowledge through time. Finally, the firm must operate subject to the transition equation for price which is dictated by the price-adjustment procedure of the regulators, namely,

$$p_i(z_i)F\{x_1[(i-1)T], x_2, x_3[(i-1)T]\} - r_2x_2 = sx_1[(i-1)T] . \tag{27}$$

This price-transition equation represents the fair-rate-of-return constraint in our model incorporating regulatory lag. It states that the regulators proceed as follows in setting the price for period i. They assume that the firm does no research and no investing after the end of period $i - 1$, so that its capital stock and stock of knowledge in period i are assumed to be $x_1[(i-1)T]$ and $x_3[(i-1)T]$, respectively. Then, the regulators calculate what price will just yield to the firm the fair rate of return in period i on its assumed capital stock for that period. In fact, equation (27) may be regarded as an implicit equation for x_2 since $x_1[(i-1)T]$, $x_3[(i-1)T]$, s and r_2 are all given at the beginning of period i. When the solution value for x_2 is obtained, p_i, the price the firm is allowed to charge in period i, follows from the inverse demand function, $p = p[F(x_1, x_2, x_3)]$.

We are still investigating the implications of this model, and others similar to it, for the issues described earlier. The methods of analysis used to examine these questions are, primarily, the standard tools the economist employs in studying problems involving optimization over time: optimal control theory, dynamic programming, and nonlinear programming. In our case, we use these techniques to characterize the optimal path for the firm maximizing the discounted present value of profit subject to the particular formulation of the regulatory constraint. For example, in the model just presented, we use these methods to study the optimal path for the firm maximizing (24) subject to (25)–(27). Then one examines the conditions characterizing this path to see whether or not the firm's use of inputs along its optimal trajectory is efficient. In particular, the efficiency or inefficiency characterizing the equilibrium point to which the firm's optimal path leads it is also examined. To investigate the effect of the length of the regulatory lag on the firm's behavior, an experiment in comparative dynamics is then performed. The parameter representing the regulators' speed of response—in the model just presented, this parameter is T—is varied, and one observes the effect of this variation on the path chosen by the firm. In particular, one performs a special comparative-statics experiment at the firm's equilibrium point to determine the effect of the change in the regulatory lag on the equilibrium stock of knowledge.

The main reason for presenting the model here is that it appears, despite its oversimplification, to be a closer representation of the regulatory process than has been available before and because it encompasses the important policy issues that were just discussed. While results are still being derived for this and similar models, it would be fair to say that the general impression suggested by the conclusions already obtained is that there is a definite role for regulatory lag to play in inducing regulated firms to undertake research, *but* that the exact relationship—even the direction of the relationship—between the length of the lag and the equilibrium stock of knowledge depends critically on the nature of the production function $F(x_1, x_2, x_3)$ and, particularly, on the way the stock of knowledge enters that function. A similar remark applies to the effects of regulatory lag on the A-J conclusions. In some models, notably the Bailey-Coleman model, the inefficiency in the capital-labor ratio declines as the lag increases. But here too there is a great deal yet to be done.

9. Evaluative comments

■ As a contribution to the theory of regulation the work of Averch and Johnson represents a significant landmark. It shows how rigorous methods can be used to draw illuminating conclusions in an area in which historical description and impressionistic discussion has until recently played a preponderant role in the literature. This is clearly to the good, and it is hard to see how this accomplishment can be questioned.

This still leaves open, however, the issue of the significance of the A-J results for policymaking. And here, it seems to us, the conclusion we should draw is rather mixed. The analysis is certainly pertinent in pointing out that rate-of-return regulation can impose its social costs in the form of input inefficiency. Unfortunately, no one has yet found a good substitute for limitation on rate of return as a device to prevent the earning of monopoly profits by public utility firms, an objective that few persons probably would be willing to abandon. Yet any constraint on the decision process which is not designed explicitly to avoid undesired interference with the allocation of resources is likely to produce distortions in their usage. Most taxes have such effects, and it is hardly surprising that a fair-rate-of-return constraint will have such consequences.

The next step, then, is to determine the most important of these undesirable effects and to consider what can be done about them. The A-J effect certainly seems to be a candidate, although we have seen that in any particular model it may or may not be present, depending on the behavioral assumptions employed. Thus, for example, in the case where the firm is a sales maximizer we have seen that excessive labor rather than capital intensiveness follows from the premises. But even in a world populated by profit-maximizers we cannot be sure that the consequences for public welfare of the A-J effect are likely to be very serious. Harberger, in a well-known article [4] has estimated, on the basis of a calculation that was admittedly crude, that the social loss through misallocation resulting from monopoly and elements related to it is surprisingly small—a rather negligible proportion of the GNP. Leibenstein [9] following this line of thought has suggested that a far more serious source of welfare loss is what he calls "X inefficiency"—cases where management, characteristically as a result of imperfect knowledge, has simply failed to organize the firm's operations in a highly efficient manner.

Regulation seems to provide particularly apt examples of this distinction. The A-J overcapitalization is an example of the inefficiencies emphasized in the more conventional analyses. But even if it occurs in practice it does not seem likely to produce effects that are very serious. This seems so, if for no other reason, because, as we have already observed, firms may have neither the extensive information nor the refined decision processes necessary to lead unerringly to the A-J input distortions.

Of course, even if the A-J overcapitalization effect is not widespread, it does not follow that rate-of-return regulation is without its dangers.[20] It has been suggested that it has sometimes been used by regulators as a means to keep inefficient firms in business, shoring them up by preventing their rivals from passing their lower costs on to the consumers. The regulator's apparent distaste for the demise of any business firm no matter how tenuous its economic justification may well contribute in no small measure to X inefficiency.

Moreover, a rate-of-return ceiling by its very nature tends to contribute another X-inefficient influence. By ruling out all profits in excess of some predetermined rate as potential fruits of monopoly, a rate-of-return constraint also precludes extraordinary rewards for extraordinary entrepreneurial accomplishment. Indeed, if such a constraint were applied with complete inflexibility it might well eliminate altogether any financial reward for efficiency and innovation. If the firm cannot hope to earn more than the ceiling, *even temporarily*, and if the demands for its products are inelastic so that it can reasonably count on that profit rate in any event, what can management hope to gain by working for increased efficiency in its operations? Fortunately, as we have indicated, regulatory lag, like the Schumpeterian innovation process, by permitting temporary rewards for efficiency can perhaps provide just the necessary incentives. This is not the place to pursue the issue further. The point is simply that while regulation may well be suspected of being the source of some non-negligible inefficiencies in the economy, it is not clear that the phenomenon encompassed by A-J analysis is the most disquieting of these.

References

1. Averch, H and Johnson, L. L. "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, Vol. 52 (December 1962), pp. 1053–69.
2. Bailey, E. E. and Coleman, R D. "A Mathematical Model for Regulatory Lag," unpublished memorandum, Bell Telephone Laboratories, October 1969.
3. Bailey, E. E. and Malone, J. C. "Resource Allocation and the Regulated Firm," *Bell Journal of Economics and Management Science*, Vol. 1, No 1 (Spring 1970), pp. 129–142.
4. Harberger, A. C. "Using the Resources at Hand More Effectively," *American Economic Review*, Vol. 49 (May 1959), pp. 134–47.
5. Kafoglis, M. Z. "Output of the Restrained Firm," *American Economic Review*, Vol. 59 (September 1969), pp. 583–89.
6. Kahn, A. E. "The Graduated Fair Return: Comment," *American Economic Review*, Vol. 58 (March 1968), pp. 170–73.
7. Klevorick, A. K. "The Graduated Fair Return: A Regulatory Proposal," *American Economic Review*, Vol. 56 (June 1966), pp. 577–84.

---

[20] On this issue see also the alternative theoretical analyses of some other problems of rate-of-return regulation in Wellisz and in Westfield.

8. ——— and Baumol, W. J. "Regulation of Public Utility Profits and Operating Efficiency," unpublished manuscript, 1968.
9. Leibenstein, H. "Allocative Efficiency vs. 'X Efficiency,'" *American Economic Review*, Vol. 56 (June 1966), pp. 392–415.
10. Takayama, A. "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, Vol. 59 (June 1969), pp. 255–60.
11. Wellisz, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *Journal of Political Economy*, Vol. 71 (February 1963), pp. 30–43.
12. Westfield, F. M. "Regulation and Conspiracy," *American Economic Review*, Vol. 55 (June 1965), pp. 424–43.
13. Wichers, C. R. "A Note on Klevorick's Regulatory Proposal," unpublished memorandum, Bell Telephone Laboratories, 1970.
14. Zajac, E. E "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model," *American Economic Review*, Vol. 60 (March 1970), pp. 117–125.

Peak-load pricing of telephone calls

S. C. Littlechild

Graduate Centre for Management Studies
Birmingham, England

Because of the complex network structure of a telephone system, it is not immediately apparent what constitutes the marginal cost of a telephone call. This paper develops a mathematical programming model based upon the maximization of producers' plus consumers' surplus subject to capacity constraints for each item of equipment at each time of day. An optimal peak-load pricing (and investment) structure results from the Kuhn-Tucker conditions: namely, that the price of a call is set equal to the sum of the marginal opportunity costs (given by the dual variables) of the capacities utilized. The model is illustrated using cost and demand data relevant to a simple network centered on Chicago, and suggests that present prices are considerably above marginal costs, especially on long-distance, high-density routes.

■ Schemes for improving the efficiency with which resources are used in the communications industry depend on the pricing of services being brought into line with marginal costs. This is not easily accomplished in practice, particularly with a telephone system, which consists of an extremely complex network structure. For several reasons, it is not immediately apparent what constitutes the marginal cost of a telephone call at any particular time. First, each item of equipment supplies capacity jointly during each period of the day. Second, the demand for a telephone call on any route is actually a joint demand for a chain of capacities. Finally, most equipment is not dedicated to a particular route but is utilized by different routes at different times.

This paper develops a mathematical programming model to determine optimal prices in this network structure. The model is based on the maximization of consumers' plus producers' surplus, subject to the constraints that for each item of equipment and in each period of the day capacity should be adequate to serve the volume of calls routed over it. The dual variables to these constraints may be interpreted as marginal opportunity costs of capacity. An optimal peak-

---

The author holds a Bachelor of Commerce degree from the University of Birmingham and the Ph D. degree from the University of Texas at Austin. He is currently Senior Research Lecturer in Economics at the Graduate Centre for Management Studies at Birmingham Research for this paper formed part of Dr. Littlechild's Ph.D dissertation, it was subsequently revised under a postdoctoral fellowship from the AT&T Company The research was also supported by Northwestern University and by the Foundation for Management Education in London.

The author wishes to thank the officers and others of the Illinois Bell Telephone Company for their cooperation in furnishing information, they are not, of course, responsible for the use made of the information He wishes particularly to mention Messrs. J. Elliot, R. Panfil, W. Sothen, and the late Mr L. Higgins. Officials of the AT&T Company provided encouragement and assistance in the revision. The author is also indebted to A. Charnes for stimulating conversations during the course of the study, to J. Hirshleifer, D. L. Kendrick, M. Marchand, and R. Turvey for written comments and discussions, to R. R. Auray and P. O Steiner, discussants of an earlier version presented at Dartmouth College, and finally to the editors of the BELL JOURNAL. A. Lent, F. Y. Phillips, and the staff of the Vogelback Computation Center at Northwestern University gave valuable assistance in the initial programming approach.

load pricing structure results from the Kuhn-Tucker conditions: namely, price of a call in any period is set equal to the sum of the marginal opportunity costs, in that period, of the items of capacity utilized.

The model is illustrated by application to a small network centered upon Chicago, using cost and demand data obtained in discussion with a local telephone company. These data suggest that present charges for calls are considerably above marginal costs, especially on long-distance, high-volume routes.

The paper is laid out as follows. Section 1 presents some general considerations pertaining to costs, prices, and revenues in the telephone system. Section 2 describes the relevant telephone technology. Section 3 formulates the model mathematically. Section 4 applies it to a simple network. Section 5 discusses the results. The appendix sets out cost and demand data.

1. Costs, prices, and revenues: general considerations

■ An excellent discussion of the principles upon which telephone charges should be set can be found in Hazlewood's 1951 paper, "Optimum Pricing as Applied to Telephone Service" [6]. These principles have recently been reiterated in the National Board for Prices and Incomes' *Report on Post Office Charges in Great Britain* [17]. Briefly, assuming that the tariff should equal marginal costs, a fixed charge per unit of time should equal the marginal cost of adding a subscriber to the system and of maintaining him there, while a call charge should equal the marginal cost of making that call.

To give an example, in Evanston, Illinois, residential subscribers currently pay $8.00 for initial connection, a fixed charge of $5.10 per month for basic service, and then a price per call that depends upon distance, duration, and time of day. However, local calls are free within Evanston, so that part of the basic service charge is a substitute for call charges.[1] Finally, non-basic items of customer equipment ("vertical services") can be obtained at additional charges.

The costs of maintaining a subscriber in the system, denoted "customer costs," are those costs

> . . . directly allocable to the individual subscriber but which are independent of the number of calls he makes. They consist of the capital cost, converted to an annual basis, of the subscriber's telephone, the line connecting him to the exchange and the terminating equipment at the exchange together with the cost of meter reading and of rendering regular accounts.[2]

The costs involved in making a call are of two types: "traffic costs" and "capacity costs." Traffic costs are those entailed by a call regardless of when it is made—principally, the cost of operators and of ticket processing. Capacity costs are the costs of providing adequate amounts of equipment in and between exchanges to meet *peak* demands on the system. Finally, there are certain amounts of "overhead costs," consisting of research and development expenditures and some administrative expenses.

A rough estimate of the magnitudes of the costs involved may be gained by examining the published annual report of the Illinois Bell Telephone Company for 1967. The first column of Table 1 presents

---

[1] In Chicago there is no free calling area, but basic service costing $5.60 includes an allowance of 80 free "message units," subsequent message units being billed at 4¾ cents On the average, 63 of the 80 units of the basic allowance are used up, so that it might be argued that the fixed charge is in fact $2 29 (= 5.60 − 63 × 4¾).

[2] [6], p. 70.

TABLE 1

INCOME STATEMENT[1]
ILLINOIS BELL TELEPHONE, 1967

COSTS (MILLIONS OF DOLLARS)	TOTAL	ADJUSTED[2]	TYPE OF COST			
			CUSTOMER	TRAFFIC	CAPACITY	OVERHEAD
TRAFFIC (OPERATORS)	68	73	11	61	0	1
COMMERCIAL	31	33	16	11	0	6
MARKETING	33	35	0	0	0	35
ACCOUNTING	26	27	14	6	0	7
LICENSE CONTRACT (A T &T.)	8	8	0	0	0	8
UNCOLLECTIBLES	4	4	2	2	0	0
OTHER OPERATING EXPENSES	29	30	0	0	0	30
PENSIONS	43					
LESS EXPENSES CHARGED CONSTRUCTION INCLUDING INTEREST	17					
MAINTENANCE	138	155	258	0	362	0
DEPRECIATION	111	111				
FEDERAL INCOME TAX	107	107				
STATE AND LOCAL TAXES AND SOCIAL SECURITY	91	91				
INTEREST	20	20				
DIVIDENDS	110	110				
RETAINED EARNINGS	26	26				
TOTALS:	830	830	301	80	370	79

TABLE 1 (CONTINUED)

REVENUES (MILLIONS OF DOLLARS)	TOTAL	TYPE OF REVENUE			
		CALL CHARGES	CUSTOMER CHARGES	CALL OR CUSTOMER CHARGES	OTHER
LOCAL SERVICE	522	115	200±	200±	
TOLL SERVICE	273	273			
MISCELLANEOUS[3]	35				35
TOTALS	830	388	200 ±	200 ±	35

[1] FORMAT OF PUBLISHED ACCOUNTS HAS BEEN SIMPLIFIED AND REARRANGED.

[2] PENSIONS ALLOCATED TO TRAFFIC, COMMERCIAL, MARKETING, ACCOUNTING AND MAINTENANCE, AND CONSTRUCTION EXPENSES RECLAIMED FROM TRAFFIC, COMMERCIAL, MARKETING, ACCOUNTING AND OTHER OPERATING EXPENSES, ON PRO RATA BASES

[3] DIRECTORY ADVERTISING AND EXTRAORDINARY ITEM (SALE OF INVESTMENT).

a simplified income statement, modified in the second column by allocating pension and construction costs. On the right-hand side an attempt is made to allocate costs among the aforementioned four categories.[3] Maintenance, depreciation, taxation, and financial costs are allocated in the same proportions as plant and equipment, valued at cost at $24 billion in 1967, of which about $10 billion was customer equipment (telephones and subscriber lines) and the balance was capacity for making calls. It appears that the costs associated with making a call amount to roughly half of the annual total, almost all of which are capacity costs rather than traffic costs. Customer costs account for most of the remainder. Overheads net of miscellaneous revenues ($= 79 - 35$) amount to less than 5% of total costs.

Of the $522 million local service revenues, $115 million comes from message units in excess of the basic allowance, and approximately half the remainder comes from optional customer equipment. The balance of some $200 million comes from basic service revenues, but, as noted in footnote 1, part of this is in fact a charge for usage. Thus, revenue from call charges is somewhere between half ($= 273 + 115 \div 830$) and three quarters ($= 273 + 115 + 200 \div 830$) of total revenues. It seems likely, then, that revenue from calls may be up to half as much again as the associated costs.

Of particular interest for present purposes is the question whether charges for peak period calls correctly reflect peak costs. If 70% of all calls are made in the daytime, or peak, period, and if off-peak prices are at about two thirds of the level of peak prices, then peak revenues amount to some 80% of call revenues ($= 0.7 \times 1 \div 0.7 \times 1 + 0.3 \times \frac{2}{3}$) whereas peak costs, consisting of capacity costs plus 70% of traffic costs, amount to nearly 95% of call costs. (If another two evening hours are included in the peak period, which now accounts for about 80% of all calls, then peak revenues constitute some 85% of total, and peak costs some 96%.) To the extent that these figures are accurate, they suggest that calls in the peak period are being subsidized by those made at off-peak hours.[4]

However, this type of approach is inadequate for developing or assessing a pricing policy on an individual call basis. A more detailed examination of telephone technology, costs, and demand is necessary.

2. Telephone technology

■ The telephone system consists of a set of local switching offices ("central offices") where calls originate and terminate, a set of tandem and toll switching offices through which calls are routed on their way to and from central offices, and trunk transmission facilities joining the offices. The lines serving individual customers are connected to the central offices, but, in the present model, individual customer equipment will be omitted. Thus we can represent the system as a network, with switching offices at the nodes and trunks as the links.

In practice, a call can be routed over any of a number of alternative paths, which saves costs by utilizing capacity more fully. For example, calls along the East Coast during the morning peak can be switched via St. Louis, where there is spare capacity because of the time difference. Although the model could include the routing problem, it will simplify the analysis considerably to assume fixed routing patterns, so that a given call, identified by its origin and destination, always traverses a given set of trunks and offices.

---

[3] These allocations and subsequent empirical calculations were made after discussion with numerous officers of the Illinois Bell Telephone Company. They should be considered as "ball-park" estimates only.

[4] It also seems that basic service is being subsidized by vertical services, for only a small proportion of the plant (hence, of customer costs) consists of vertical service equipment, whereas over half of customer revenues are from vertical services. This question is beyond the scope of the present paper, however.

Since the demand for calls at any time is a stochastic function, the concept of capacity requires elaboration. The theoretical maximum capacity of a telephone channel (a voice-grade circuit)[5] is 3600 call seconds per hour, written 36 ccs (also denoted one erlang). That is, the channel can handle one call lasting an hour, or 360 calls lasting 10 seconds each, and so on. But it is rare that a call begins precisely as an earlier one ends. In practice any given channel stands idle for long periods and at other times is busy when a second call is attempted. The term "effective capacity" may be introduced. Effective capacity depends upon the stochastic character of the traffic (the more erratic it is, the lower is the mean number of ccs a channel can handle) and upon the number of channels available (two channels can handle more than twice as much traffic as one can). Effective capacity is defined for a given quality of service, by which is meant the probability of there being a free channel at any time.

A combination of theoretical and empirical methods has shown that the traffic on a typical intrastate toll route allows a mean demand of about 25 ccs per channel at a 97% quality of service. On a more heavily used interstate route, effective capacity is about 30 ccs per channel, whereas on an intracity route it varies from 30 ccs down to 10 ccs.

Ideally, the optimal quality of service should be determined within the model.[6] But because of the difficulties of evaluating costs of delayed calls so as to make them comparable with equipment costs, it would be misleading to embody such costs in the present model. Accordingly, the standard of 97% quality of service used by the Bell System has been adopted.[7] Of course, it is quite possible to use techniques of sensitivity analysis in order to explore the tradeoff between quality of service and cost.

Similar considerations of capacity apply to the switching equipment within a central office, where customers' lines are connected to each other and to incoming and outgoing trunk circuits. That is, the effective capacity of such equipment is determined by the quality of service specified. Again, the Bell System standard has been used.

Switching offices also contain varying amounts of equipment that is not directly associated with the paths over which calls are routed. Such equipment is at a minimum in a step-by-step office without common control, where the switches themselves are controlled by dial pulses from the customer's telephone. At the opposite extreme are modern electronic switching offices, which contain large amounts of electronic storage and control facilities that are summoned into service for particular calls as needed.

---

[5] Strictly speaking, of a pair of channels, or circuits, to permit two-way conversations

[6] See Marchand [15]. In particular, M Marchand has pointed out that optimal resource allocation may well require significant variations in the quality of service for different types of calls at different times of day. Minasian [16] observes that off-peak customers do in fact enjoy service of higher quality.

[7] Actually, the concept of quality of service in a network is rather more complicated. If a call requires a chain of, say, four items of equipment, each of which has a 97% probability of being free, then the probability of completing a call is roughly $(0.97)^4 \simeq 0.88$. But if there are two alternative routes at each link of the chain, the probability of traversing each link is raised to $1 - 0.03^2 \simeq 0.999$ and the probability of completing the call is $(0.999)^4 \simeq 0.996$. These complications will be ignored here. On the relationship between quality of service and capacity, see [17], Appendix J.

In a crossbar system (assumed in the present model) control equipment consists of markers and registers. These set up paths through the switching equipment for incoming and outgoing calls. The control equipment drops off as soon as its function is performed; it is not required for the duration of the call. The work of a switchboard operator has the same characteristic. These considerations provide the justification (from the cost point of view) for charging a higher rate for the first minute, say, than for succeeding minutes. However, examination of cost figures (see Appendix, part 4a) suggests that, for a local call of average duration, switching and control costs are in the approximate ratio four to one, and that for toll calls the ratio is very much higher. Accordingly, control equipment will not be distinguished separately, its cost being included with that of the switching equipment. This means that the cost of a short call is slightly understated relative to that of a long one.

3. Mathematical formulation of the model

■ **The basic model.** Problems of peak-load pricing arise when capacity is supplied jointly to all periods of the day and demand is not homogeneous over the day. Charging a common price leads to capacity standing idle in some periods even though its marginal opportunity cost is zero. Steiner [18] formulated (and illustrated geometrically) an approach to electricity pricing based upon the maximization of consumers' and producers' surplus. The solution involved charging a price in each period such that demand in each period equalled capacity, thereby flattening out the load curve (with the provision that price should not be below marginal energy cost). As Hirshleifer [7] pointed out, the resulting (market-clearing) prices, net of marginal energy costs, represent marginal opportunity costs or values of capacity in the respective periods (for "opportunity" should be understood as including alternative consumption as well as production opportunities). Capacity should then be chosen so that its marginal value, as given by summing over all periods of the day the marginal opportunity costs, is just equal to its marginal purchase cost.[8]

Williamson [21] suggested a reformulation of the model as a constrained maximization problem, with the Lagrangian multipliers yielding marginal opportunity costs of capacity. In fact, the peak-load pricing problem is one of a class of joint cost problems which can be conveniently treated using mathematical programming techniques, and in which optimal prices appear as sums of dual variables [13]. Such a programming approach extends naturally to the case of joint demand [12]. It is thereby straightforward to handle the problems peculiar to a telephone system, namely, that a telephone call is a joint demand requiring a whole chain of items of equipment, each of which is used by calls on several different routes and in different proportions depending upon the time of day.[9]

---

[8] Similar results had been obtained by Boiteux [3] and Houthakker [9], using different approaches. For more recent treatments, see Dreze [5] and Marchand [15].

[9] Kalaba and Juncosa [11] presented in programming format two cost-minimizing models of the telephone system, designed to optimize routing of calls for given facilities and given demand, and to optimize design of facilities for given demand. Tulkens [19] presents a cost-minimizing analysis of some analogous problems in the postal service.

One further problem arises in application. Previous models have been quite static, assuming long-run or fully adjusted demand and cost curves. But in practice full adjustment is never achieved in a dynamic economy. In general, a multi-period framework is required. But it has been shown elsewhere that if it is optimal to maintain capacity fully utilized, then an equivalent static model may be obtained by correctly defining the annual equivalents of the installation or capital costs.[10] Annual cost may be taken as a proportion of capital cost, where this proportion equals the sum of the rates of physical depreciation, obsolescence, maintenance, cost of capital (including allowed rate of return and taxation) etc., and implies zero excess profits. Thus defined, annual cost is also equal to the cost of bringing investment forward by one year.

☐ **Analytic formulation of the model.** Let there be calls over $j = 1, \ldots, n$ different routes during $k = 1, \ldots, q$ different periods of the day, requiring capacity on $i = 1, \ldots, m$ items (or groups of items) of equipment. On the assumption of a fixed routing pattern, for each route j the set A_j of items of equipment used by every call on that route can be identified. Each set A_j will include the originating central-office equipment, the receiving central-office equipment, toll or tandem equipment, and certain specified links or channels. Similarly, for each item i of equipment, the set B_i of routes that use equipment can also be identified.

There are two types of variables, both required to be non-negative.

$x_j(k)$ = the mean volume of (three-minute) calls per hour on route j during period k.

y_i = effective capacity (in number of calls per hour) provided on equipment i.

The periods k should be chosen such that demand within each one is relatively homogeneous. But since metering and billing considerations generally dictate that only a few periods be distinguished, mean demand may vary significantly from hour to hour within each period. The constraints must ensure that capacity is adequate to meet not just mean demand, but rather maximum demand, in each period. Accordingly, define the following intraperiod peak factors,

$h_j(k)$ = ratio of maximum demand to mean demand for calls on route j in period k to mean demand on that route in that period,

and write the constraints for each item of equipment in each period of the day

$$\sum_{j \in B_i} h_j(k)x_j(k) \leqslant y_i, \quad i = 1, \ldots, m;\ k = 1, \ldots, q. \tag{1}$$

In general, the demand for calls on any route in any period is a function of prices obtaining on that route in all periods. This interdependence causes certain theoretical and practical difficulties.[11]

---

[10] See Littlechild [12], following Arrow [1]. This result requires the purchase cost per unit of new capacity to be constant with respect to the number of units purchased, i.e., no indivisibilities or economies of scale, on which see section 4.

[11] Dreze (in [5], p. 17) observed that a solution could be obtained by successive approximations, but that uniqueness had not yet been proved. Hotelling (in [8], p. 247) argued that although the value of the solution was in general dependent upon the path of integration, the same value would be obtained if the integrability conditions were satisfied. In the present context, these require that the partials of demand in period k [k'] with respect to price in period k' [k] be equal.

However, there is some evidence that although cross-elasticities are significant on higher-priced very long-distance (926–3000 miles) routes, they are not significant for the shorter routes which will be embodied in the present application.[12] Accordingly, it will be assumed that demand in any period depends upon price in that period alone.[13] Let the (inverse) demand functions be given by

$p_j(x_j(k), k)$ = price (in cents per call) at which mean demand per hour on route j in period k is $x_j(k)$, and let

$t(k)$ = length (in hours) of period k, where $\sum_k t(k) = 24$.

It will be argued in the next section that, for pricing purposes, constant marginal costs may be assumed, so let

c_j = marginal traffic cost (in cents per call) on route j, and

β_i = marginal cost of equipment i (in cents per day per three minute call capacity).

The maximand is the sum of producers' and consumers' surplus given by the integrals of the demand curves less costs, or[14]

$$\sum_j \sum_k t(k)\left[\int x_j(k)\, p_j(x_j(k), k)\, dx_j(k) - c_j x_j(k)\right] - \sum_i \beta_i y_i . \tag{2}$$

☐ **Analysis of the model.** If the demand curves are downward sloping, hence the objective function concave, the Kuhn-Tucker conditions for the system (1), (2), are necessary and sufficient for optimality. Introducing dual variables $t(k)u_i(k)$, the Kuhn-Tucker conditions are, upon dividing through by $t(k)$ where appropriate,

$$\sum_{j \in B_i} h_j(k)x_j(k) \leqslant y_i , \qquad u_i(k) \geq 0 \tag{3}$$

$$p_j(x_j(k), k) - c_j \leqslant h_j(k) \qquad \sum_{i \in A_j} u_i(k),\ x_j(k) \geq 0 \tag{4}$$

$$\sum_k t(k)u_i(k) \leqslant \beta_i \qquad y_i \geq 0 . \tag{5}$$

Complementary slackness conditions are understood to hold between each pair of inequalities. That is, if one of the pair holds with strict inequality, then the other must hold with equality.

The term $u_i(k)$ may be interpreted as the (hourly) marginal opportunity cost (in cents per call) of capacity i in period k.[15] If an item of capacity is not fully utilized at any time, its marginal opportunity cost is zero

$$\sum_{j \in B_i} h_j(k)x_j(k) < y_i \Rightarrow u_i(k) = 0 . \tag{3'}$$

---

[12] See Auray [2], exhibits 3, 4, and 5 The most noticeable effect of evening price reductions is a shift from person-to-person to station-to-station calls in the same period

[13] It is suggested elsewhere [12] that it may be possible to avoid interdependencies in the demand functions by an appropriate redefinition of variables and the introduction of constraints.

[14] It will not cause any confusion to use the upper limit of integration as the variable of integration also The lower limit may be taken as any small positive number in the event that $p(\cdot)$ is not defined at zero. Alternatively, the present level of output may be used as the lower limit. The magnitude of (the increase in) total consumer surplus is not of concern here. Prices will depend solely upon the levels of demand in the neighborhood of the optimum and not upon what could be extorted for the first few units of output. Recall in what follows that

$$d/dx \int^x f(x')dx' = f(x)$$

[15] See Hirshleifer [7], p 458

Price of a call in any period is set equal to its marginal traffic cost plus the sum of the marginal opportunity costs at that time[16] of the capacities it requires, allowing for the spare capacity necessitated by the peak factor

$$x_j(k) > 0 \Rightarrow p_j(k) = c_j + h_j(k) \sum_{i \in A_j} u_i(k) . \qquad (4')$$

Equipment is purchased to the point where the marginal value of the capacity it provides in all periods equals its marginal cost

$$y_i > 0 \Rightarrow \sum_k t(k) u_i(k) = \beta_i . \qquad (5')$$

At optimality, one might think of the marginal cost of capacity in any period k' as the marginal purchase cost of additional equipment less the marginal opportunity value of the capacity it provides in all other periods, for

$$t(k') u_i(k') = \beta - \sum_{k \neq k'} t(k) u_i(k) . \qquad (5'')$$

4. An application

■ The model was applied to the simple three-route network shown in Figure 1 (page 200), illustrating a local call (within Chicago, Ill.: $j = 1$), an intrastate toll call (Chicago to Peoria, Ill.: $j = 2$), and an interstate toll call (Chicago to New York, N.Y.: $j = 3$).

The day was divided into four periods, corresponding to those obtaining for pricing purposes in Illinois in 1967 (to which year all data refer), except that the hours 6:00 a.m. to 8:00 a.m. were included in the day period because the hourly volume of calls then is of an order of magnitude (i.e., about ten times) higher than in other after-midnight hours. The periods are:

Day ($k = 1$):	6:00 a.m. to 6:00 p.m.,	$t(1) = 12$
Evening ($k = 2$):	6:00 p.m. to 8:00 p.m.,	$t(2) = 2$
Night ($k = 3$):	8:00 p.m. to 12:00 p.m.,	$t(3) = 4$
After-midnight ($k = 4$):	12:00 p.m. to 6:00 a.m.,	$t(4) = 6$.

Tables 2 and 3 (page 201) set out the data which were used to fit linear demand and cost functions. Further details are given in the appendix. For simplicity, the example is limited to direct-dialed station-to-station (DDD) calls in a completely automated switching system. Operators are nevertheless required for assistance, information, and intercept work, and the costs of these services were averaged over the number of calls made, as were the costs of uncollectibles, of handling complaints, and of ticket processing. These traffic costs c_j total 5.3 cents per toll call and 0.6 cents per local call, as set out in Table 3(a). It should be noted that, strictly speaking, they are average traffic costs, and marginal traffic costs may be lower.[17]

Each route requires several types of equipment, some of which are used in common, and which for present purposes may be grouped

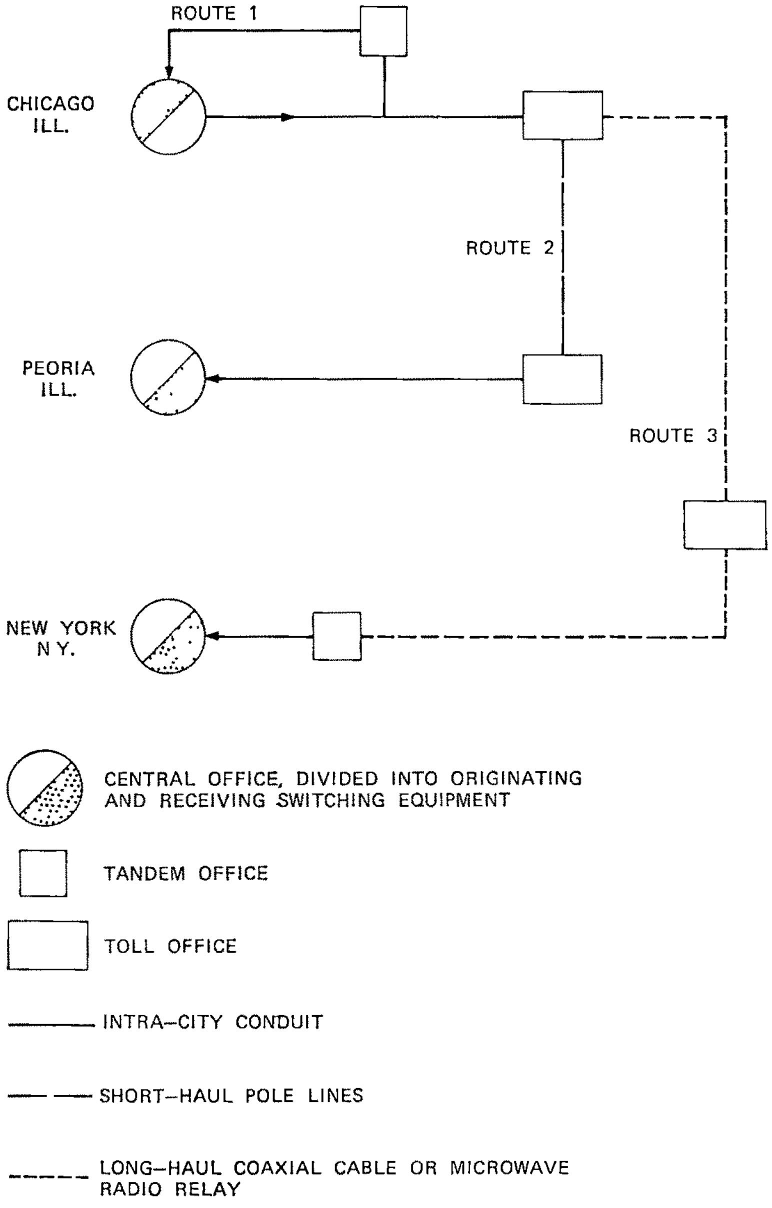
---

[16] If, instead, it is desired to charge a price per call independent of the time of day, introduce the demand function $p(x)$ in the objective function and the terms $a(k)x$ in the constraints, where $a(k)$ is the proportion of calls made per hour in period k The price of a call will then be given by $\Sigma\, a(k)u_i(k)$, that is, the average marginal cost Similarly, in lieu of charging separately for (say) local calls, one could determine a fixed charge equal to the average marginal cost of making the average number of calls This leads into questions of whether "the product" is a telephone call or the ability to make one, which are beyond the scope of this paper

[17] M Wilk observed that much traffic-recording equipment is determined by *maximum* calling rate, in which case it is an item of capacity cost rather than traffic cost.

FIGURE 1

EXAMPLE NETWORK


ROUTE 1
CHICAGO ILL.
ROUTE 2
PEORIA ILL.
ROUTE 3
NEW YORK N Y.
CENTRAL OFFICE, DIVIDED INTO ORIGINATING AND RECEIVING SWITCHING EQUIPMENT
TANDEM OFFICE
TOLL OFFICE
INTRA-CITY CONDUIT
SHORT-HAUL POLE LINES
LONG-HAUL COAXIAL CABLE OR MICROWAVE RADIO RELAY

into five "items." Items $i = 1, 2, 3$ are used by, respectively, routes $j = 1, 2, 3$ alone, item $i = 4$ is used by toll routes 2 and 3, and item $i = 5$ is used by all routes.[18] Table 3(b) sets out the marginal capacity costs β_i for these items.

Certain comments are in order regarding the derivation of these cost coefficients. There are three significant sources of economies of scale in the telephone system, which were dealt with as follows. First, as remarked earlier, because of the stochastic nature of the traffic, an increase in volume on any trunk group can be handled by a less than proportionate increase in the number of circuits there while maintaining the same quality of service. It was assumed that, within

---

[18] Thus, in terms of the notation of the previous section, $B_1 = \{1\}$, $B_2 = \{2\}$, $B_3 = \{3\}$, $B_4 = \{2,3\}$ and $B_5 = \{1,2,3\}$. Hence $A_1 = \{1,5\}$, $A_2 = \{2,4,5\}$ and $A_3 = \{3,4,5\}$.

TABLE 2

INITIAL PRICE AND USAGE DATA

ROUTE j	PERIOD k	INITIAL PRICE[1] $p_j^o(k)$	INITIAL USAGE[2] $x_j^o(k)$	INTRA-PERIOD PEAK FACTOR[3] $h_j(k)$	ELASTICITY $\epsilon_j(k)$
INTRA-CHICAGO $j = 1$	DAY	5	445	1.6	−0.1
	EVE	5	486	1.2	−0.1
	NIGHT	5	210	1 8	−0.1
	AFTER M.	5	10	1.8	−0.1
CHICAGO-PEORIA $j = 2$	DAY	65	73	1.4	−0 15
	EVE.	50	78	1.0	−0 37
	NIGHT	40	27	2.1	−0 36
	AFTER M.	40	3	1.8	−0 30
CHICAGO-NEW YORK $j = 3$	DAY	140	2570	1.4	−0 23
	EVE.	100	3873	1.0	−0 57
	NIGHT	70	1940	1 8	−0.57
	AFTER M	70	148	1.8	−0.56

[1] IN CENTS PER THREE MINUTE CALL.

[2] EQUIVALENT NUMBER OF THREE MINUTE CALLS PER HOUR ON AN AVERAGE BUSINESS DAY. INTRA-CHICAGO FIGURES IN THOUSANDS.

[3] RATIO OF MAXIMUM NUMBER OF CALLS IN ANY HALF HOUR IN EACH PERIOD TO AVERAGE NUMBER OF CALLS PER HALF HOUR IN THAT PERIOD.

TABLE 3

COST DATA

(a) TRAFFIC COSTS	
ROUTE j	COST[1] c_j
INTRA-CHICAGO, $j = 1$	0 6
CHICAGO-PEORIA, $j = 2$	5 4
CHICAGO-NEW YORK, $j = 3$	5.4

(b) CAPACITY COSTS	
ITEM i	COST[2] β_i
TANDEM OFFICE AND INCOMING SWITCHING EQUIPMENT IN CHICAGO, $i = 1$	5.3
CABLE FROM CHICAGO TO PEORIA, TOLL OFFICE AND INCOMING SWITCHING EQUIPMENT IN PEORIA, $i = 2$	37 8
COAXIAL CABLE FROM CHICAGO TO NEW YORK, TWO TOLL OFFICES AND INCOMING SWITCHING EQUIPMENT IN NEW YORK, $i = 3$	75 2
TOLL OFFICE IN CHICAGO, $i = 4$	8 2
ORIGINATING SWITCHING EQUIPMENT AND INTRA-CITY CONDUIT IN CHICAGO, $i = 5$	11.2

1 IN CENTS PER CALL.

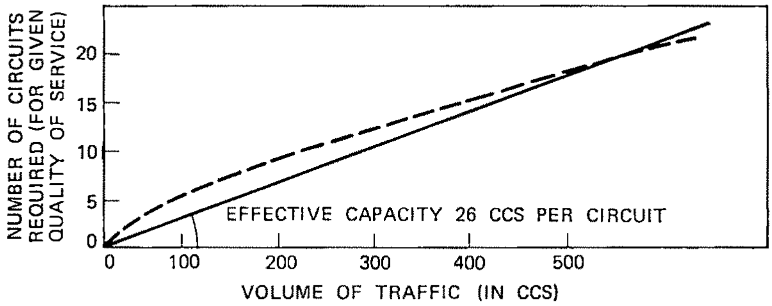
2 IN CENTS PER DAY PER (THREE MINUTE) CALL CAPACITY.

the range of traffic likely to obtain on any trunk group, effective capacity per circuit could be approximated by a constant equal to the average capacity per circuit at the expected volume. Thus, in Figure 2(a), at a traffic volume of 500 ccs, effective circuit capacity would be taken to be 25 ccs (which compares with a marginal circuit capacity at that volume of 29 ccs). Evidently, effective capacity will differ according to the volume and type of traffic in the trunk group.

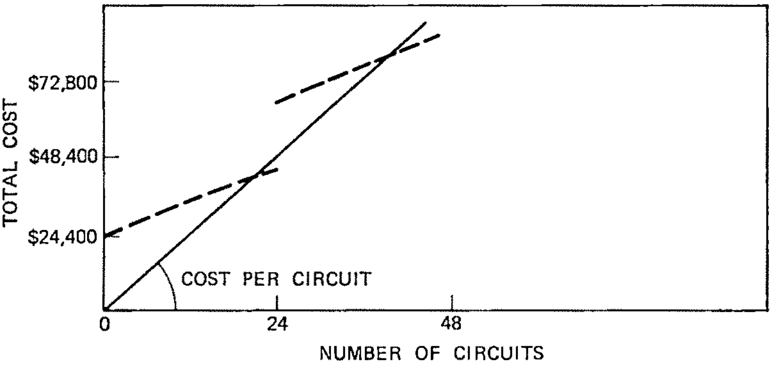
Second, for each type of trunk transmission facility, there are cost components incurred per system-mile, regardless of the number of circuits, as well as cost components incurred per circuit-mile (as Table 4 shows). However, such "overhead" components are incurred quite regularly as traffic increases because of the limited circuit

FIGURE 2

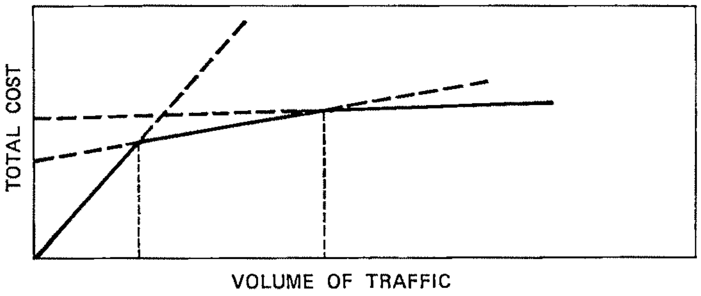
THREE SOURCES OF ECONOMIES OF SCALE



(a) EFFECTIVE CCS CAPACITY PER CIRCUIT



(b) COST PER CIRCUIT (POLE WIRE)



(c) CHOICE OF APPROPRIATE SYSTEM

– – – – DENOTES ACTUAL COST FUNCTIONS

——— DENOTES APPROXIMATION

TABLE 4

INSTALLATION COSTS OF TRANSMISSION FACILITIES

	CONDUIT (IV)	POLE LINES (V)	COAXIAL CABLE (VI)	RADIO–RELAY
(1) INSTALLATION COST PER CIRCUIT–MILE			$1.59	$3.00
(2) INSTALLATION COST PER CIRCUIT		$1,000	$1,242	$1,412
(3) TYPICAL LENGTH OF FACILITY (MILES)		100	1000	1000
(4) SECOND COMPONENT PER CIRCUIT–MILE [(2) – (3)]		$10 00	$1.24	$1.41
(5) INSTALLATION COST PER SYSTEM–MILE	$55,250	$244	$43,600	$15,800
(6) MAXIMUM NUMBER OF CIRCUITS PER FACILITY	1000	24	7700	6000
(7) TYPICAL CAPACITY UTILIZATION ("FILL")	0.85	0 90	0.80	0.80
(8) THIRD COMPONENT PER CIRCUIT–MILE [(5) ÷ (6) x (7)]	$65.00	$11 30	$7.08	$3 29
(9) FIXED COST OF INSTALLATION PER SYSTEM		$10,000		
(10) FOURTH COMPONENT PER CIRCUIT–MILE [(9)÷(3) x (6) x (7)]		$4 63		
(11) TOTAL COST PER CIRCUIT–MILE [(1) + (4) + (8) + (10)]	$65 00	$25.93	$9 91	$7.70
(12) HOURLY EFFECTIVE CAPACITY PER CIRCUIT (CCS)	20	25	30	30
(13) CAPITAL COST PER CCS PER MILE [(11) ÷ (12)]	$3.25	$1.04	$0 33	$0.26

NOTE COSTS OF MICROWAVE RADIO RELAY ARE PROVIDED FOR COMPARISON WITH COAXIAL CABLE AS AN ALTERNATIVE LONG–HAUL TRANSMISSION FACILITY.

capacity of each type of transmission facility. Further, because of the interdependencies and alternate routing possibilities in the network, which enable a higher and more even level of capacity utilization to be obtained,[19] it seemed appropriate for pricing purposes to regard a whole transmission facility or office as the marginal unit[20]—that is, to treat the problem as one of "lumpiness" which should be evened out. Marginal costs per circuit-mile were then derived on the basis of typical capacity utilization, or "fill." For example, to install 100 miles of pole line costs $24,400 plus $1,000 per circuit up to 24 circuits (see Figure 2(b) and Table 4). If such a system typically has a 90-percent fill, then cost per circuit was taken as $1,000 + $24,400/ 24 × 0.9) = $1,000 + 1,130) = $2,130). A similar argument was applied to office equipment.

Finally, there are economies of scale with respect to size and distance of trunk group. Table 4 shows that microwave radio relay systems or coaxial cable systems carrying thousands of circuits over long distances are much cheaper per circuit-mile than short-haul pole lines carrying, say, two dozen circuits over a hundred miles. In the present study it was assumed that the most economical type of transmission facility for each trunk group could be predetermined (Figure 2(c)).

These assumptions combine to yield marginal capacity cost equal to average capacity cost for each item of equipment. This is analytically convenient, it ensures that total costs are just recouped, and it reflects economies of scale by as fine a partition of trunk-group types as desired. However, two difficulties encountered in a dynamic

---

[19] Alternate routing possibilities make a network cheaper than a collection of separate routes, but it is not clear whether such economies increase with size nor how far they are offset by switching costs.

[20] Houthakker [9] made a similar argument for the electricity network.

FIGURE 3

CONSTRAINT STRUCTURE OF EXAMPLE

VARIABLES z_j^k OUTPUTS												VARIABLES y_i CAPACITIES					RIGHT HAND SIDE
z_1^1	z_1^2	z_1^3	z_1^4	z_2^1	z_2^2	z_2^3	z_2^4	z_3^1	z_3^2	z_3^3	z_3^4	y_1	y_2	y_3	y_4	y_5	
1												−1					⩽ 0
	1											−1					⩽ 0
		1										−1					⩽ 0
			1									−1					⩽ 0
				1									−1				⩽ 0
					1								−1				⩽ 0
						1							−1				⩽ 0
							1						−1				⩽ 0
								1						−1			⩽ 0
									1					−1			⩽ 0
										1				−1			⩽ 0
											1			−1			⩽ 0
				1				1							−1		⩽ 0
					1				1						−1		⩽ 0
						1				1					−1		⩽ 0
							1				1				−1		⩽ 0
1				1				1								−1	⩽ 0
	1				1				1							−1	⩽ 0
		1				1				1						−1	⩽ 0
			1				1				1					−1	⩽ 0

system must be noted. First, the appropriate type of facility for a growing trunk group will change over time, and prices should be reduced accordingly, but the earlier-installed, more expensive equipment remains. Second, in anticipation of growth, facilities may be installed in excess of present requirements because of setup costs of installation (c.f. Manne [14]). Alternatively, a different type of facility might be installed, appropriate to some future level of traffic. In either case there is a conflict between short-run marginal-cost pricing and recouping costs. There is, of course, a long-standing dispute between the advocates of marginal-cost pricing (e.g. Hotelling [8] and Vickrey [20]) and the skeptics (e.g., Coase [4] and Wiseman [22]), and the issue will not be treated further here.

5. Results

■ Applying the transformation

$$z_j{}^k = h_j(k)x_j(k) \qquad (6)$$

to the constraint set (1) yields the structure set out in Figure 3 (while maintaining the separability and concavity of the objective function). This structure suggests that the general problem offers considerable scope for computational short cuts such as the decomposition methods of Dantzig and Wolfe or the mixing routines of Charnes and Cooper, so that very large networks may thereby be tackled.

The present example is simple enough to be solved on a desk calculator, and it is instructive to examine the prices obtained at successive stages of solution, as set out in Table 5. Suppose first that the network were to be made entirely separable by route—that

TABLE 5

OPTIMAL PRICES AND USAGE

ROUTE j	PERIOD k	(i) PRICES WITH ITEMS i=4,5 DEDICATED[1]	(ii) PRICES WITH ITEM i=4 DEDICATED[1]	(iii) OPTIMAL PRICES[1] $p_j^*(k)$	(iv) OPTIMAL USAGE[2] $x_j^*(k)$
1	1	2.8	2.8	2.8	465
	2	0.6	0.6	0.6	529
	3	0.6	0.6	0.6	229
	4	0.6	0.6	0.6	10
2	1	12.0	12.0	11.1	82
	2	5.4	5.4	9.4	101
	3	5.4	5.4	5.4	35
	4	5.4	5.4	5.4	4
3	1	5.4	6.7	6.7	3163
	2	43.7	40.7	40.7	5062
	3	9.8	8.5	8.5	2852
	4	5.4	5.4	5.4	225

[1] IN CENTS PER THREE MINUTE CALL

[2] EQUIVALENT NUMBER OF THREE MINUTE CALLS PER HOUR ON AN AVERAGE BUSINESS DAY. INTRA-CHICAGO FIGURES IN THOUSANDS

is, presently-shared originating central-office switching equipment and conduit transmission facilities in Chicago (item $i = 5$) and the Chicago toll office (item $i = 4$) would instead be divided and "dedicated" to a particular route, and each route would provide all its own equipment. The ensuing prices are set out in column (i). It turns out that the Chicago and Peoria routes are fully utilized in the daytime, the New York route in the evening and the night.

Suppose now that the originating switching and transmission equipment in Chicago ($i = 5$) is shared. The prices of evening and night-time calls to New York are reduced (column ii) because they do not have to cover originating switching costs (since other routes have such capacity spare then), but the price of New York daytime calls increases because this period now has to provide its own originating switching capacity.

Finally, when Chicago toll office facilities ($i = 4$) are shared between the New York and Peoria routes, the latter uses the former's excess capacity in the daytime, thereby reducing Peoria daytime prices (column iii), but toll capacity still has to be purchased for Peoria calls in other periods, and the evening bears the cost.

It becomes apparent that the technological interchangeability of capacity in the telephone network not only reduces the total amount of equipment necessary for any given set of demands, thereby lowering the average level of prices, but it may also have the effect of lessening the differentials between prices in different periods of the day. This occurs when peaks are experienced for different routes at different times. One might, therefore, find prices to be much more even over the day in a larger, more realistic model, especially with alternate routing allowed.

From Table 2, the present average revenue or price per call (weighted by usage at each time of day) is 5 cents for the local call, 60 cents for the Peoria call, and $1.20 for the New York call. From Table 5, the optimal average revenues or prices are about 2 cents, 10 cents, and 12 cents, respectively. That is, the optimal average revenue for the local call is about half the present revenue, for the intrastate call about one sixth, and for the interstate call about one tenth.

The order of magnitude difference between optimal and present prices is considerably greater than the prediction of the Income Statement calculations of Section 1 (that present call revenues might be at most half again as much as costs). One possible explanation is that the costs of providing free local calls in fact outweighs the basic service charge in many (especially rural) areas, with the deficit being covered by profits on toll calls and other (urban) local calls. Another possible explanation is that the telephone system, having grown up piecemeal, does not consist of the most efficient types of equipment which would be purchased now, and which were used as a basis for the cost estimates.

Given that prices are above marginal costs, the price ratio should be higher on the routes with less elastic demand (whether the company is maximizing profit or minimizing loss of consumer surplus). In fact, precisely the opposite is at present the case. This may reflect the pressure of State regulatory commissions which assume themselves to be judged by their ability to keep down local and intrastate rates. It also seems to have been widely accepted that local calls are a social necessity in a way that toll calls are not.

Although New York is nearly ten times farther away from Chicago than is Peoria, the optimal average price is a mere twenty percent higher. The reason for this, as the Appendix shows, is that coaxial cable on a high-volume route costs only about one third (and radio relay about one quarter) as much per circuit-mile as does cable on a low-density route. The present average price to New York, which is twice that to Peoria, only partly reflects those cost differences. In fact, the policy of uniform pricing by distance has led to the emergence of competition on high-density routes, to which the response has been the filing of tariffs for point-to-point pricing on a number of such routes.[21]

Turn now to the optimal structure of prices by time of day. For local calls within Chicago, the daytime period bears the whole of the capacity charge, with calls in other periods being priced simply at their traffic cost. On the Peoria route, the daytime and evening periods share the capacity cost. On the New York route, it turns out that capacity costs are borne almost entirely by the evening calls, with some support from daytime and night-time calls. On all routes, calls after midnight are priced at their traffic cost alone.

The price structure on the first two routes is quite plausible, but that on the New York route seems to call into question the assumption of independence of demand between periods. When the daytime price drops to one sixth of that obtaining in the evening, it does not seem reasonable to maintain that demand will not shift. Two observations are in order.

First, the volume of calls on the New York route is at present higher in the evening than in the daytime (and in fact there is evidence that the Bell System peak has shifted to the evening). Further, if the elasticity estimates of Table 2 are correct, this would continue to be the case if the present evening price of \$1.00 was raised to \$1.50, i.e., above the daytime price of \$1.40. Peak prices in the evening are therefore not implausible.[22]

---

[21] Perhaps an alternative tariff could be constructed to reflect costs more accurately than at present, while preserving simplicity of rate structure and guarding against arbitrariness, by specifying a "national grid" of high-density routes with prices to be charged at a given rate to and from the grid and at a given but lower rate on it.

Second, a linear approximation of short-run demand cannot be expected to give adequate results when price is reduced to five percent of its original level, as is the case in the daytime on the New York route. At such low prices one might expect new sources of demand to develop in the long run, thereby reducing price differences within the day, and it is clearly necessary that the demand curves reflect this. It was hoped that the present exercise would identify certain fruitful areas for further theoretical and empirical research. Evidently long-run demand is one such area.

Appendix[23]

1. Elasticities

Studies carried out by the Long Lines Department of A.T.&T. suggest daytime elasticities on 500–700 mile routes of about −0.2 and −0.3 for business and residential subscribers, respectively, and on 100-mile routes elasticities of −0.1 and −0.2, respectively. Night-time elasticities are approximately double. Elasticities for local calls have not been found significantly different from zero, and −0.1 is commonly used. The parameters $\epsilon_j(k)$ were obtained by weighting the above figures by the estimated proportions of business and residence originated calls in each period.[24]

2. Demand functions

For each route j and period k, linear demand functions $p = a - bx$ were fitted, where $b = (- p^0/\epsilon x^0$ and $a = p^0(1 - 1/\epsilon)$. Elasticity is of course defined by $\epsilon = dx/dp \cdot p/x$, evaluated at initial (1967) observations p^0, x^0. Parameters a and b are given by

j	k	$a_j(k)$	$b_j(k)$
1	1	55	0.1124
	2	55	0.1029
	3	55	0.2380
	4	55	5.2083
2	1	501	5.952
	2	185	1.739
	3	152	4.193
	4	172	40.404
3	1	756	0.2369
	2	270	0.0453
	3	189	0.0633
	4	196	0.8475

---

[22] Incidentally, it does not seem to have been observed in the literature that when demand curves cross (as linear ones often will), then whichever period is thought of intuitively as the peak (i.e., which bears the greatest proportion of marginal capacity cost) will depend upon the amount of capacity installed, and hence upon the level of capacity cost in relation to demand.

[23] Further notes on the calculations contained in the appendix can be obtained from the author upon request.

[24] Recent econometric studies by the G.P.O. in Great Britain (reported in [10]) found price elasticities of demand of − 0.55 for trunk calls (average distance about 55 miles) and − 0.12 for local calls

3. Traffic costs

Component	Cost per call in cents Toll	 Local
1. Operators (assistance, information and intercept)	0.43	0.43
2. Commercial (contact work, e.g., complaints)	1.95	0.10
3. Accounting (ticket processing: recording and pricing)	2.63	0.02
4. Uncollectibles	0.35	0.02
Total	5.36	0.57

4. Capacity costs

Six basic types of equipment are identified; their capital or installation costs are denoted $G_{\mathrm{I}}, \ldots, G_{\mathrm{VI}}$. An item of equipment $i (i = 1, \ldots, 5)$ consists of one or more of these types, weighted appropriately; its capital cost is denoted g_i. The latter are reduced to annual costs by a factor θ, thence to the daily costs β_i.

(a) Central Office

Installed cost = \$1,007,000 + \$20 per originating ccs per hour
+ \$19 per incoming ccs per hour
+ \$24 per originating attempt per hour
+ \$10 per incoming attempt per hour.

(In addition, there is a cost of \$23 per line, belonging to customer cost.) The overhead of \$1,007,000, mainly land, buildings and power equipment, was arbitrarily allocated to the originating switching equipment, assumed to be operating at 80% of its 27,000 ccs per hour capacity. Control equipment costs were added to switching equipment costs on the basis of 1.3 attempts per completed call and and average duration of call of 3 ccs. Thus, capital cost per ccs of originating switching equipment

$$G_{\mathrm{I}} = 20 + \left(\frac{24 \times 1.3}{3} + \frac{1{,}007{,}000}{72{,}000 \times 0.8}\right) = \$47.9,$$

and capital cost per ccs of incoming switching equipment

$$G_{\mathrm{II}} = 19 + \left(\frac{10 \times 1.3}{3}\right) = \$23.3\,.$$

(b) Toll (or Tandem) Switching Office

Installed cost = \$2,000,000 + \$1,000 per trunk.
Assuming the toll office operates at 60% of its 20,000 trunks capacity, and a trunk has an effective capacity of 25 ccs, capital cost per ccs of toll office capacity

$$G_{\mathrm{III}} = \frac{1{,}000}{25} + \left(\frac{2{,}000{,}000}{20{,}000 \times 0.6}\right) = \$46.7\,.$$

(c) Trunk Transmission Facilities (see Table 4)

For each of three types (IV, V, VI) of transmission facilities, installation cost may be expressed as a linear function of number of circuit-miles (1), number of circuits (2), number of miles (5), plus an overhead component (9). The last three were reduced to costs per circuit-mile (4), (8), (10) by assumptions on average length of transmission facility (3), maximum number of circuits per facility (6), and typical capacity utilization (7). Total cost per circuit mile (11) was obtained by summing (1), (4), (8) and (10) and dividing by effective capacity per circuit (12) to yield the following capital costs per ccs-mile (13)

$$G_{\mathrm{IV}} = \$3.25, \quad G_{\mathrm{V}} = \$1.04, \quad G_{\mathrm{VI}} = \$0.33.$$

Local calls require incoming switching equipment (II) in Chicago, and about 15% of them use a tandem office (III) there; hence cost of item $i = 1$ used by Chicago local calls alone is

$$g_1 = G_{\mathrm{II}} + 0.15G_{\mathrm{III}} = \$30.305.$$

The Peoria route requires 125 miles of pole lines (V), a further toll office (III), 5 miles of conduit (IV), and incoming switching equipment (II) in Peoria:

$$g_2 = 125G_{\mathrm{V}} + G_{\mathrm{III}} + 5G_{\mathrm{IV}} + G_{\mathrm{II}} = \$216.25.$$

The New York route requires 900 miles of coaxial cable (VI), 2 further toll offices (III), 4 miles of conduit (IV), and incoming switching equipment (II) in New York:

$$g_3 = 900G_{\mathrm{VI}} + 2G_{\mathrm{III}} + 5G_{\mathrm{IV}} + G_{\mathrm{II}} = \$429.95.$$

Both toll routes share a Chicago toll office (III):

$$g_4 = G_{\mathrm{III}} = \$46.7.$$

All routes require orginating switching equipment (I) in Chicago and an average of about 5 miles of conduit (IV) in Chicago to either a receiving office or a toll office:

$$g_5 = G_{\mathrm{I}} + 5G_{\mathrm{IV}} = \$64.15.$$

The annualization factor θ is given by the sum of the rates of maintenance (8.15%), plus property taxes (4.79%), physical depreciation and obsolescence (5.84%), and cost of capital i.e., interest, income tax, dividends and retained earnings (13.84%), totalling 32.6%. Strictly speaking, θ_i should be calculated for each item i; the above rates are averages given by the annual Income Statement figures (Table 1) as percentages of written down book value of plant ($19 billion).

The g_i are capital costs expressed in dollars per ccs; assuming 330 working days per year, equivalent daily costs β_i in cents per three-minute (= 1.8 ccs) call are given by

$$\beta_i = \left(\frac{100 \times 1.8 \times 0.326}{330}\right) g_i = 0.175 g_i,$$

as set out in Table 3(b).

References

1. ARROW, K. J. "Optimal Capital Policy, the Cost of Capital, and Myopic Decision Rules," *Annals of the Institute of Statistical Mathematics*, Vol. 16, Nos. 1-2 (Tokyo: 1964), pp. 21-30.
2. AURAY, R. R. "Discussion of 'Marginal Cost Pricing of Telephone Calls' by S. C. Littlechild," a paper presented at the Public Utilities Seminar, Dartmouth College, September, 1969.

3. Boiteux, M. "La Tarification de Demandes en Point: Application de la Théorie de la Vente au Coût Marginal," *Revue Générale de l'Electricité*, Vol. 58 (August 1949), pp 321–40. Translated as "Peak-Load Pricing," *Journal of Business*, Vol. 33 (April 1960), pp. 157–79.
4. Coase, R. H. "The Marginal Cost Controversy," *Economica*, NS (August 1946), pp. 169–82
5. Drèze, J. H. "Some Post-War Contributions of French Economists to Theory and Public Policy with Special Emphasis on Problems of Resource Allocation," *American Economic Review*, Vol 54, No. 4, Part 2 (June 1964).
6. Hazlewood, A. "Optimum Pricing as Applied to Telephone Service," *Review of Economic Studies*, Vol 18, No. 2 (1950–51), pp. 67–78
7. Hirshleifer, J. "Peak Loads and Efficient Pricing: Comment," *Quarterly Journal of Economics*, Vol 72, No. 3 (August 1958), pp. 451–62.
8. Hotelling, H. "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, Vol. 6, No. 3 (July 1938), pp. 242–69.
9. Houthakker, H. S. "Electricity Tariffs in Theory and Practice," *Economic Journal*, Vol. 61, No. 1 (March 1951), pp. 1–25
10. International Telegraph & Telephone Consultative Committee. *Economic Studies in Telecommunications (1964–68).* Geneva: International Telecommunications Union, July 1, 1968.
11. Kalaba, R. E. and Juncosa, M. L. "Optimal Design and Utilization of Communications Networks," *Management Science*, Vol 3, No. 1 (October 1956), pp. 33–44
12 Littlechild, S. C. "On the Representation of Cost and Demand Interdependencies in Public Utility Pricing," Discussion Paper No. 12, Graduate Centre for Management Studies, Birmingham, England (January 1970).
13. ———. "Marginal Cost Pricing with Joint Costs," *Economic Journal* (June 1970).
14. Manne, A. S "Capacity Expansion and Probabilistic Growth," *Econometrica*, Vol. 29, No. 4 (1961).
15 Marchand, M. "Economic Principles and Telephone Rates," Working Paper No. 251, Center for Research in Management Science, University of California, Berkeley (April 1968).
16. Minasian, J. R. "Ambiguities in Theory of Peak-Load Pricing and Application of Theory of Queues," *Land Economics*, Vol. 42, No. 3 (August 1966), pp. 355–62.
17. National Board for Prices and Incomes. *Post Office Charges.* Report No. 58, Cmnd. 3574 (March 1968), pp. 260–73.
18. Steiner, P. O. "Peak Loads and Efficient Pricing," *Quarterly Journal of Economics*, Vol 71, No. 4 (November 1957), pp. 585–610.
19 Tulkens, H. *Programming Analysis of the Postal Service.* Louvain: Librairie Universitaire, 1968.
20 Vickrey, W. "Some Objections to Marginal Cost Pricing," *Journal of Political Economy*, Vol. 56 (1948).
21. Williamson, O E. "Peak-Load Pricing and Optimal Capacity Under Indivisibility Constraints," *American Economic Review*, Vol. 56, No. 4 (September 1966), pp. 810–27.
22 Wiseman, J. "The Theory of Public Utility Price—an Empty Box," *Oxford Economic Papers*, (1957), pp. 56–74.

Rate of return and business risk

Paul H. Cootner and Daniel M. Holland

Sloan School of Management
Massachusetts Institute of Technology

This paper reports in succinct form the findings of a study undertaken a number of years ago that sought to measure the relation between risk and rate of return and, thus, to determine an empirical basis for implementing several Supreme Court decisions (Bluefield Waterworks *Case, 1923, for one) that a public utility is entitled to earnings sufficient to permit its rate of return on invested capital to be similar to those in "other business undertakings which are attended by corresponding risks and uncertainties." Defining "business" risk as functionally related to the variability of earnings, a number of hypotheses concerning rate of return and risk are tested statistically, and a reasonable and significant association is discovered between them for both a sample of industries and individual companies. The statistical model, however, explains only about a quarter of the variability in rates of return among industries and firms, suggesting that it has not captured some important determinants of business risk. The final section examines our results in the light of recent developments in the theory of financial risk.*

Introduction

■ This paper reports the major results of a study completed several years ago. In some respects it shows its age. Intervening years have been characterized by expansion and improvement in data bases useful for studies of this sort and, perhaps more important, the development of new theory relevant to the questions considered here. During this time, our ideas and findings have been elaborated upon by some and disputed by others, in both academic and regulatory contexts, despite the unavailability of any published version of the study.

At the request of the editors of the BELL JOURNAL, we have agreed to resurrect these eight-year-old ideas. At the same time, we would like to seize the opportunity to place them in the perspective of a decade of tremendous theoretical progress in dealing with the theory of risk and painfully slow acceptance of new developments in the regulatory area.

When we started our research in 1961, there were virtually no published quantitative studies of the relation between risk and return

---

Dr. Cootner has been Professor of Finance at M.I.T. since 1966 and is currently Visiting Professor of Finance at Stanford University. He holds the B.S. and M.S. degrees from the University of Florida (1949 and '50, respectively) and the Ph.D. in Economics from M.I.T. (1953). He taught at Brown and Northwestern prior to working as a Research Associate at Resources for the Future from 1956 to 1959. He joined the M.I.T. faculty in 1959 as an Assistant Professor.

Dr. Holland holds the A. B. and Ph.D. degrees from Columbia University (1941 and 1951, respectively). He has taught at Columbia, New York University, and, since 1958, at the Sloan School of Management at M.I.T., where he is presently Professor of Finance. He has been on the research staff of the National Bureau of Economic Research since 1951 and Editor of the *National Tax Journal* since 1966.

This paper is drawn from a non-technical summary of a larger unpublished study by the authors, *Report on Study of Risk and Rate of Return* (sponsored by the American Telephone and Telegraph Company), Massachusetts Institute of Technology DSR Project 9565, March 1963. The authors are grateful to Dr. W. F. Sharpe for his help in the preparation of this paper.

in business enterprise, especially in the field of regulated industries. Regulatory commissions and legal precedent eschewed the use of capital market data: indeed, to this day the "imbedded cost of debt" is considered to be the relevant item for rate-of-return purposes rather than the current cost, and for the last five years at least, hearing examiners have lectured economists that current rates were irrelevant because they were at historical peaks and "had to come down." In that sort of context we set out to develop a measure of risk which could be estimated from accounting data.

Even if we had not placed that restriction upon ourselves, it is quite unlikely that we would have discovered the very important theorems about financial risk later developed in the work of Sharpe, Lintner, Fama, and Mossin. We both now agree that in perfectly organized rational markets those theorems should describe the required trade-offs between return and risk. Furthermore, since we feel that actual capital markets are close to perfect, these predicted trade-offs should be close to the relationship actually demanded in the market place. In point of fact, however, the empirical evidence is not as unequivocal as it might be, and since the question is, essentially, an empirical one, there is still interest in these hypotheses on business risk. Despite our appreciation of the beautiful results of the recent theory of financial risk, we do not feel especially apologetic about our older results. Within the class of risk measures utilizing accounting data we consider our measures superior to their competitors.

In the first four sections we set forth a brief summary of our 1962 monograph. In the concluding section we re-examine these results in light of recent progress in the theory of financial risk. Finally, in the appendix we reproduce some results in which we sought to examine the generality of our findings by estimating the relationships we posited and tested for book value data for a comparable set of financial data. The original monograph has a more detailed explanation of our methods and a more complete analysis of issues and findings.

Risk and rate of return

■ In the *Bluefield Waterworks* case (1923) the Supreme Court held:

> A public utility is entitled to such rates as will permit it to earn a return on the value of the property which it employs for the convenience of the public equal to that generally being made at the same time and in the same general part of the country on investments in other business undertakings which are attended by corresponding risks and uncertainties; but it has no constitutional right to profits such as are realized or anticipated in highly profitable enterprises or speculative ventures.

Two decades later, in its decision in the *Hope Natural Gas* case (1944) the Court repeated this point of view.

> From the investor or company point of view it is important that there be enough revenue not only for operating expenses but also for capital costs of the business. These include service on the debt and dividends on the stock. By that standard the return to the equity owner should be commensurate with returns on investment in other enterprises having corresponding risks.

In general, then, through the thinking of the Supreme Court has run a thread to this effect: a regulated utility should be permitted to earn at a rate equal to that experienced by firms in (unregulated) industries characterized by comparable or corresponding risks.

The comparable earnings rule as it has developed in these two decisions leaves open not only the question—What is risk and how is the rate of return related to it?—but also the problem—Whose risk and return are relevant?

To amplify on this latter point: in 1923 the Court spoke about the enterprise and the risks and uncertainties associated with it; therefore it was addressing itself to the relation between the assets the firm administered and the income stream it generated from their administration. The Court seemed to be concerned with the business's risk, not with the risk undertaken and return obtained by a particular group within the firm. Risk was considered to be a characteristic of the business, and it is properly descriptive, therefore, to call this business risk. In the 1944 version of the comparable earnings rule, however, equity owners are singled out for direct attention; the Court specifically dwelt on the return to them.

Our efforts, however, have concentrated on the conception of risk in the earlier and more general Court formulation. We deal, then, with business risk instead of financial risk, addressing our inquiry to an attempt at an empirical answer to the questions raised above: "What is risk and how is the rate of return related to it?"

Once we commit ourselves to accounting data, defining rate of return poses problems of measurement but not of concept. But risk means different things to different men, and it is hard to define simply. We think it captures the essence of the problem to define situations as risky whenever there is a probability of an outcome different from the expected one. Risk is present whenever one is less than certain about the outcome of an action. A risky situation is characterized by a probability distribution of possible outcomes rather than a single, certain result. Quantitative measures of risk, then, can be developed by reference to various features of the probability distribution.

For managerial decisions, the relevant expected value of an action is not the possible monetary returns weighted by the probability of their occurrence, but the utilities of possible monetary returns weighted by the probability of their occurrence. In a risky situation, managers maximize, not expected money income but expected utility, and the heart of the process that represents business reaction to risk is the differential "satisfaction" associated with additional amounts of income.

This directs our attention, in particular, to two features of monetary returns—the probability distribution of outcomes and the marginal utility of income. Neither is easy to measure. As far as the latter goes, we simply reason from intuition, introspection, and the fact that managers in the usual case take a "trusteeship" attitude toward the assets they administer. The utility function they consider, therefore, is that of a risk averter, whether it is their own personal attitude or that of the shareholders.[1] As to risk itself, the thing which is associated with the probability distribution of outcomes, we note the obvious, that it is usually difficult, if not impossible, to obtain the whole distribution. We must be satisfied, therefore, with one or another feature of the distribution of possible outcomes. In the main, we will be interested in the dispersion of the distribution as measured by the standard deviation, which we take to be an efficient summary measure. However, we recognize the potential importance of other features of the distribution and have experimented with skewness as well.

---

[1] The modern theory of finance assumes that the relevant utility function is that of the shareholders. Our formulation is consistent with this view but does not exclude the possibility that managers consider only their own utility functions.

In a nutshell our hypothesis is this: there is a positive association between business risk and rate of return. The greater the risk, the higher the average rate of return.

Measuring rate of return

■ We wish to concentrate on business risk; thus we need a measure of the rate of return on the company's assets. This is by no means easy to obtain, since both income and asset values are, to a major extent, artifacts of accounting conventions. For example, due to changes in the depreciation laws since 1954, a dollar of profit in the years since means something different than it did before 1954. And the differential degree of application of the laws by various companies may lead to further distortions.

In spite of these and other drawbacks, we have used the traditional measure of rate of return on assets—net income after taxes plus interest, all divided by total capitalization (measured by book value). All information was obtained from the Studley-Shupert Company, Inc., a Boston-based investment counseling firm which supplied the statistical information to the American Telephone and Telegraph Company on a fee basis. AT&T had the data keypunched on cards and supplied them to us in that form. The sample included 315 companies, which were grouped into 39 industrial classifications for portions of the analysis.

The body of data from which our sample was drawn consisted of the balance sheet and income statement information for 388 of the larger companies in the United States from 1946 through 1960. Included were all the companies used in the Dow-Jones Industrial index (except for AT&T) and forty-six of the fifty largest manufacturing firms as of December 31, 1960. For our study, 73 of the 388 companies were excluded either because they fell in industries (SIC classification) which were represented by less than four companies in the Studley-Shupert data or because they could not reasonably be classified in a single industry. Within each of the 39 industry groups used in our study, when ascertainable (which could be done for all but a few) the companies included accounted for at least 80 percent of the industry's total assets.

Industry rate of return as a function of within-industry dispersion

■ The first attempt to quantify the relationship between risk and rate of return was a very simple exploratory model designed to explain how *industry* rates of return might vary with risk. The hypothesis tested was that the dispersion of company rates of return around the average rate of return of the industry to which they belong is an indication of the riskiness of an investment in that industry. Since the standard deviation of such rates of return indicated the likelihood that a company would fare differently from the industry average, we would expect that if executives were risk-averters large standard deviations would require high average rates of return to attract investment. Our

TABLE 1

INDUSTRY RATE OF RETURN REGRESSION EQUATIONS

YEAR	REGRESSION EQUATION	CORRELATION COEFFICIENT	YEAR	REGRESSION EQUATION	CORRELATION COEFFICIENT
ALL YEARS*	$I = 0.935\,x_1 + 8.18$ (0.230) (0.34)	R = 0.550	1953*	$I = 0.682\,x_1 + 7.63$ (0.236) (0.33)	R = 0.424
1946	$I = 0.151\,x_1 + 11.44$ (0.076) (0.88)	R = 0.087	1954	$I = 0.244\,x_1 + 8.50$ (0.263) (0.50)	R = 0.149
1947	$I = 0.611\,x_1 + 11.35$ (0.286) (0.72)	R = 0.327	1955*	$I = 1.282\,x_1 + 6.23$ (0.171) (0.33)	R = 0.771
1948*	$I = 0.780\,x_1 + 10.35$ (0.234) (0.56)	R = 0.476	1956*	$I = 0.744\,x_1 + 7.70$ (0.211) (0.25)	R = 0.503
1949*	$I = 0.578\,x_1 + 9.31$ (0.189) (0.53)	R = 0.445	1957	$I = 0.676\,x_1 + 7.59$ (0.273) (0.49)	R = 0.372
1950*	$I = 1.065\,x_1 + 9.21$ (0.167) (0.52)	R = 0.718	1958	$I = 0.491\,x_1 + 7.12$ (0.274) (0.49)	R = 0.279
1951*	$I = 0.724\,x_1 + 8.53$ (0.221) (0.22)	R = 0.469	1959	$I = 0.234\,x_1 + 9.12$ (0.190) (0.46)	R = 0.196
1952	$I = 0.323\,x_1 + 8.69$ (0.244) (0.38)	R = 0.210	1960	$I = 0.278\,x_1 + 9.87$ (0.139) (0.56)	R = 0.309

* INDICATES A STATISTICALLY SIGNIFICANT REGRESSION COEFFICIENT AT THE 1 PERCENT LEVEL. FOR TESTING SIGNIFICANCE 38 WAS THE DEGREE OF FREEDOM FOR EVERY EQUATION.

I = UNWEIGHTED AVERAGE OF COMPANY RATES OF RETURN FOR EACH INDUSTRY.

x_1 = THE SAMPLE STANDARD DEVIATION OF COMPANY RATES OF RETURN AROUND UNWEIGHTED INDUSTRY AVERAGE.

model was a simple linear one, of the form

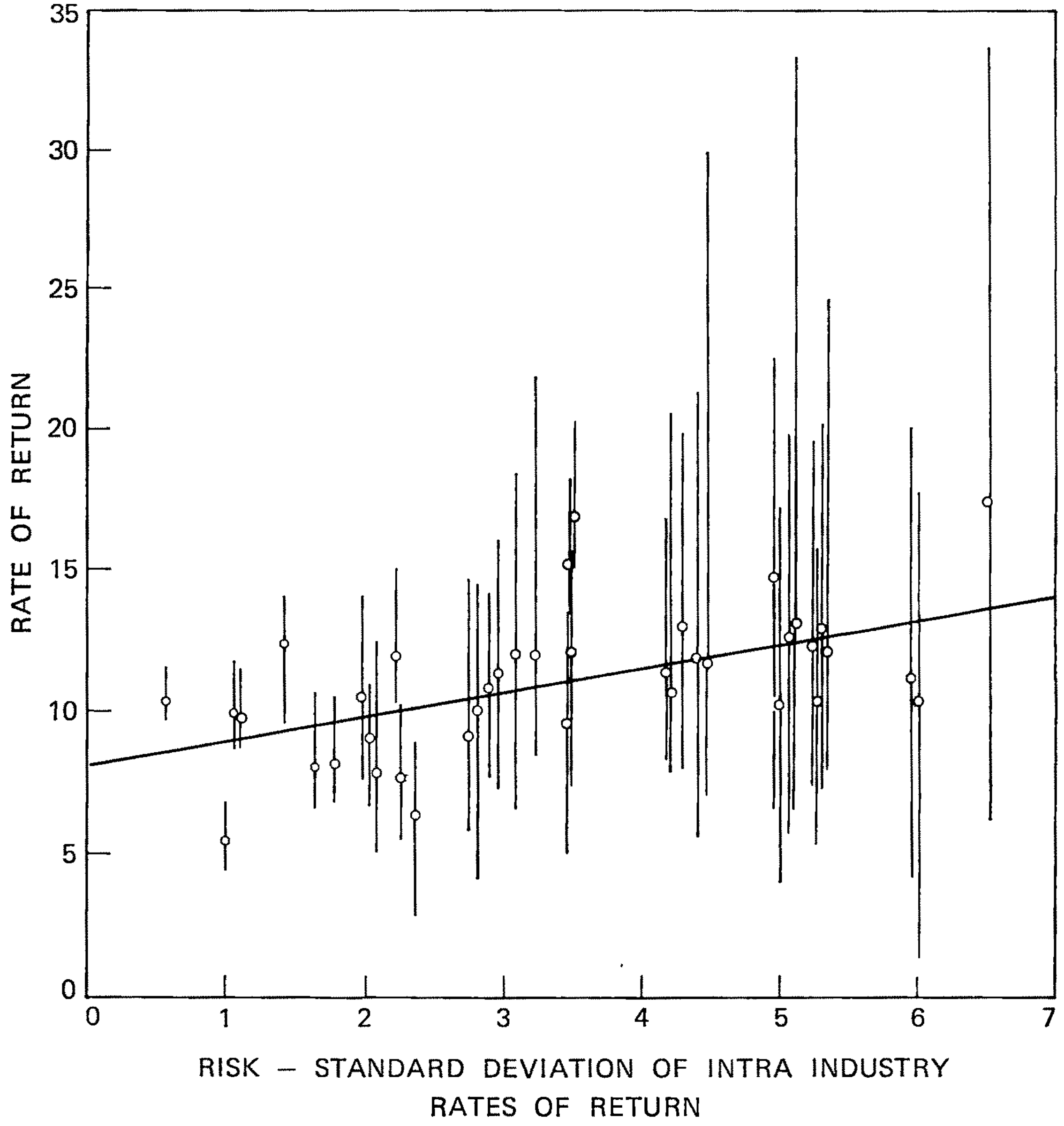
$$I = ax_1 + b\,,$$

where I is the industry average rate of return on capitalization (return is taken as income [after taxes] plus interest payments) and x_1 is the standard deviation of company rates of return around the industry average. Implicit in our model is the assumption that realized rates of return are unbiased estimates of expected returns.

The test of the model using average company rates of return for 39 industries for the entire postwar period (through 1960) indicates a positive relationship between rate of return and this measure of risk with a high degree of statistical significance (see Table 1). First of all, the sign of the coefficient of x_1 is positive, so the higher is x_1 (our measure of risk), the higher is the predicted rate of return. The equation suggests that a 1-percent increase in the dispersion is associated with almost a 1-percent increase (0.935 percent) in average industry rate of return (see the first line of Table 1). The constant term, 1.18, implies that an industry without any risk of the kind measured by x_1 would earn at a rate of 8.18 percent. This certainly seems too high (for the period covered by the study) for a completely "riskless" rate of return, even allowing for the fact that industrial investments are likely to be less easily marketable than, say, government securities. This, however, is just another way of emphasizing that 8.18 is simply the predicted rate of return for investments without this particular kind of risk. As shown below, eliminating other kinds of risk tends to reduce the "risk-free" rate of return, and

FIGURE 1

INDUSTRY RATE OF RETURN ON BOOK CAPITAL AS RELATED TO INDUSTRY RISK



eliminating risks that we have not dealt with might reduce it still further.[2]

Figure 1 depicts the relationship. The values of the variable, x_1, are pictured on the horizontal scale; the variable, I, on the vertical. The dots represent the various industry rates of return and standard deviations. The vertical lines indicate the range of rates of return within the industries represented by the dots.

---

[2] But since the constant term, b, is the repository for all we have failed to explain, as long as factors other than risk account for variations in rates of return among industries or companies, exhaustion of all possible risk variables will bring the constant term down to a "riskless" rate of return, but not necessarily down to, say, the rate on government bonds.

The correlation coefficient for this regression is 0.550; stated another way, 30 percent of the variability of industry rate of return is explained by the dispersion of company rates of return. It is well to note that the influence of this one variable upon rate of return may be overwhelmed in any particular case by the variables which explain the other 70 percent of the variability of rate of return. However, the relationship is significant—statistical tests indicate that it is very unlikely to have occurred by chance.

To test the consistency of the relationship we found for the data summed over a span of years, we computed the same regression for each of the individual years in the postwar period. The results are summarized in Table 1.

In general, the annual figures show a high degree of consistency with each other and with the all-years regression. All but one of the individual years show a positive relationship between rate of return and this measure of risk, and in seven of the fifteen regressions the relationships are statistically significant at the 1-percent level (less than one chance in one hundred of being accidental) and ten of the coefficients of x_1 are within two standard errors of the coefficient for the all-years regression. Even the constant terms make a certain amount of common sense since they are all very significant and decline from the high levels which existed in the early post-war years. It is interesting to note that while corporate rates of return continued to decline from 1955 to 1960, the constant term in the regression equations tended to rise. This may mean that in the eyes of investors, business conditions became more risky from the point of view of risks other than those measured in this model—for example: increased risk of government intervention, or of foreign competition, or of a slower overall rate of economic growth, or of changes in the rate of inflation. In any event, the results from the annual regressions are by no means implausible.

Company rates of return

■ Our next step was to develop for individual companies the same type of regression equation that we used for industry rate of return. But in this case, of course, it was individual company rate of return we sought to explain, and we used a number of independent variables jointly for this purpose. The regression equations are shown in Table 2, the results in Table 3.

The most striking result is that the more complex formulations lead to correlations which are little, if any, higher than the results obtained in the first attempt, i.e., the regression that used the dispersion of industry rates of return. The next most striking result is that only variables x_1 (the variability of company return around the industry mean return) and x_2 (the variability of company return around its own mean) contribute much to the correlations.

The first point may seem a little surprising. After all, the industry approach using variable x_1 produced correlation coefficients of 0.55 and now we have added other explanatory variables as well. The answer is that it is much harder to explain differences in company rates of return than it is to explain differences in industry rate of return. This can be shown by the following example. Suppose we could explain industry rate of return perfectly. Then, unless all the companies had the same rate of return, we would find that the

TABLE 2

GENERAL FORM OF THE COMPANY RATE OF RETURN REGRESSIONS

(a) $Y = a_1 x_1 + c_1$

(b) $Y = a_1 x_1 + a_2 x_2 + c_2$

(c) $Y = a_1 x_1 + a_2 x_2 + a_3 x_3 + c_3$

(d) $Y = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + c_4$

(e) $Y = a_1 x_1 + a_4 x_4 + c_5$

Y = THE AVERAGE RATE OF RETURN ON CAPITALIZATION FOR A COMPANY FOR THE POST–WAR PERIOD, 1946–60

x_1 = THE STANDARD DEVIATION OF THE COMPANY'S RATES OF RETURN OVER THE POST–WAR PERIOD AROUND THE MEAN OF THE INDUSTRY TO WHICH THE COMPANY BELONGS FOR THE SAME PERIOD.

x_2 = THE STANDARD DEVIATION OF THE ANNUAL RATES OF RETURN FOR THE COMPANY AROUND ITS POST–WAR AVERAGE.

x_3 = THE SKEWNESS OF THE COMPANY'S ANNUAL RATES OF RETURN.

x_4 = THE STANDARD DEVIATION OF ANNUAL CHANGES IN THE COMPANY'S RATE OF RETURN AROUND THE MEAN POST–WAR CHANGE.

formula would make mistakes in predicting individual company rates of return since it would predict the same rate for all companies in the industry.

In point of fact, the correlation of company rate of return with variable x_1 (the dispersion of company rate of return around its industry rate of return) is 0.300, but because we are now working with 315 companies instead of 39 industries, the results are even more statistically significant than before. By adding variable x_2 (the dispersion of company rate of return over the postwar years), we raise the correlation coefficient to 0.502, or roughly the same level we achieved for the industry predictions. Actually the second variable is more useful in explaining the rate of return than x_1, since x_2's simple correlation with rate of return is 0.437. If we had computed the regressions in the reverse order, i.e., first found the relationship between *Y* and x_2 and *then* between *Y* and both x_2 and x_1, we would have found that the former was 0.437 and the latter was 0.502. In other words, if *Y* is correlated with x_1 with a coefficient of 0.300,

TABLE 3

COMPANY RATE OF RETURN REGRESSION EQUATIONS

	MULTIPLE CORRELATION COEFFICIENT
(a) $Y = 0\ 955\ x_1 + 8\ 20$ (0 171) (0.235	R = 0.300
(b) $Y = 0\ 788\ x_1 + 0\ 944\ x_2 + 5.31$ (0 157) (0 115) (0 203)	R = 0 502
(c) $Y = 0\ 724\ x_1 + 0.917\ x_2 + 11.56\ x_3 + 4\ 59$ (0 157) (0.114) (4 41)	R = 0 517
(d) $Y = 0\ 732\ x_1 + 1\ 063\ x_2 + 12\ 56\ x_3 - 0\ 202\ x_4 + 4\ 63$ (0 157) (0 179) (4.51) (0 191)	R = 0 520
(e) $Y = 0\ 812\ x_1 + 0\ 724\ x_4 + 6.21$	R = 0.420

NOTE: THE NUMBERS IN PARENTHESES ARE THE STANDARD ERRORS OF THE ESTIMATED VALUES OF THE REGRESSION COEFFICIENTS then adding x_2 would add 0.202 points to our correlation, but if we correlated Y and x_2 (correlation 0.437), then adding x_1 to the relation contributes only 0.065 (0.502 − 0.437) to the correlation. Dealing with both variables instead of either one adds relatively little to the explanatory power because both measures of dispersion tend to measure the same thing to some extent. This is even more true of the variables x_3 and x_4, which explain almost nothing about rate of return that is not already explained by variables x_1 and x_2. In fact x_4 and x_2 are so closely related that the introduction of x_4 serves merely to affect the previous coefficient of x_2.

Since the regression of Y on x_1 and x_2 gives us our most useful results, let us examine it more closely. In the previous section the relationship between industry rate (I) of return and x_1 was shown to be:

$$\begin{array}{llll} I = & 0.935x_1 + & 8.18 & R = 0.55. \\ & (0.230) & (0.340) & \end{array}$$

That is, for each 1-percent increase in x_1 there was a 0.935-percent increase in the required rate of return. Correlating Y (company rate of return) with x_1 the relation becomes:

$$\begin{array}{llll} Y = & 0.955x_1 + & 8.20 & R = 0.30\,. \\ & (0.171) & (0.235) & \end{array}$$

That is, the coefficients are little changed, the lower value for R merely indicating that less of the variation is explained.

If, however, we relate Y to both x_1 *and* x_2 we get the equation

$$\begin{array}{lllll} Y = & 0.788x_1 + & 0.944x_2 + & 5.31 & R = 0.50\,. \\ & (0.157) & (0.115) & (0.203) & \end{array}$$

Note now that the coefficient of x_1 has dropped so that the industry dispersion no longer explains as big a portion of the company rate of return, although comparing it with its standard error (in parentheses below it) we see it is still very significant in the statistical sense.[3] Actually this change is to be expected, since x_2 is a factor which is much more specific to the company, is of the same order of magnitude, and overlaps x_1 somewhat, and which should, therefore, account for a portion of Y.

Looked at another way, we can say that a 1-percent change in x_1 will increase Y by 0.955 percent if we specify nothing about x_2. But x_1 and x_2 are not independent of one another, so that if x_1 increases, x_2 is likely to increase as well, though not as much. What our second equation tells us is that if x_2 is held constant and only x_1 increases by 1 percent, then Y will increase by only 0.788 percent instead of by the 0.955 percent it would increase if x_2 were not held constant. Similarly, if x_1 were held constant and x_2 were increased by 1 percent, Y would increase by 0.944 percent.

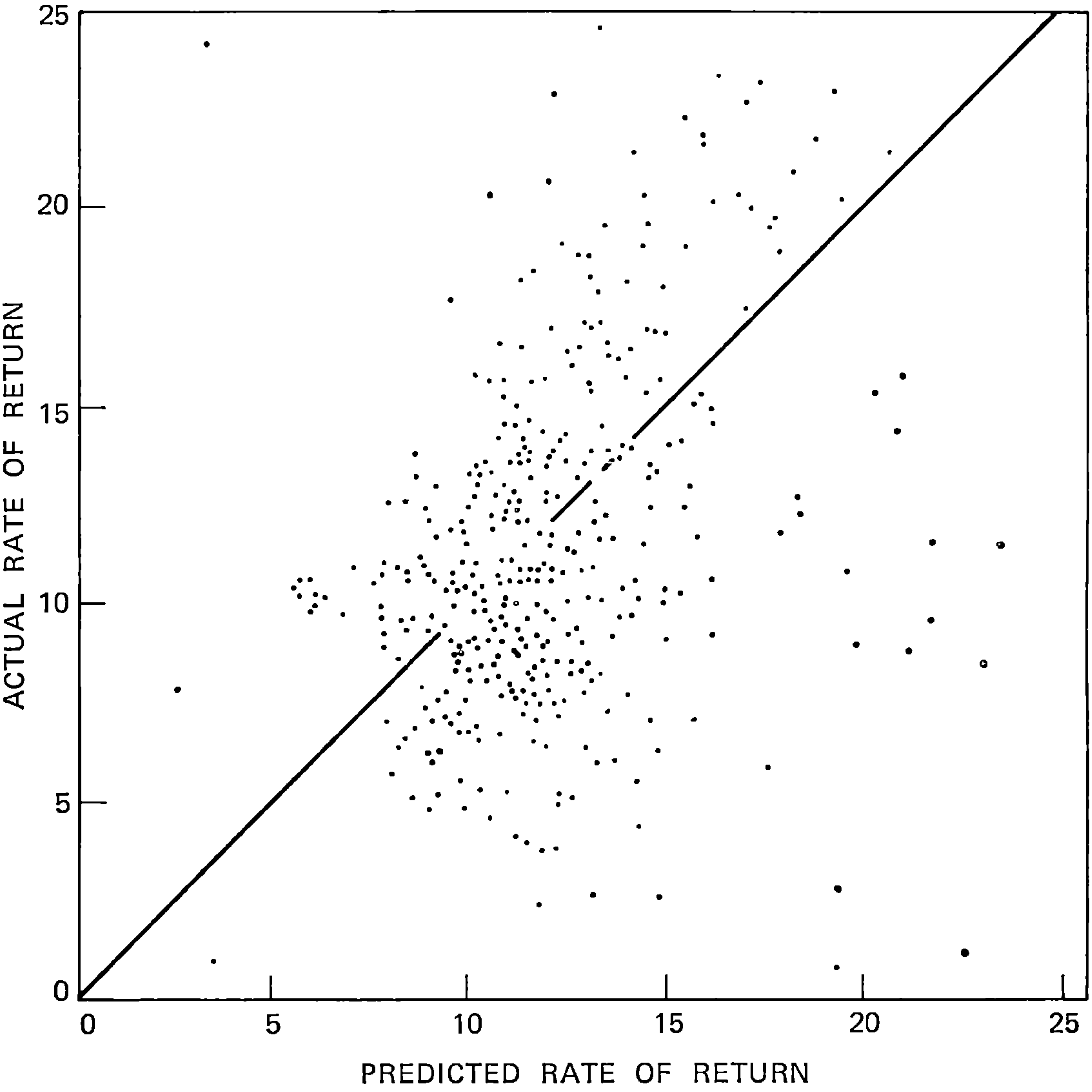
The other major change resulting from accounting for these two kinds of business risk is that the apparent "risk-free" rate of return on long-term industrial investment falls to 5.31 percent from 8.20 percent and remains highly significant. That is, this is the rate of

---

[3] Specifically, there is only one chance in a hundred that the "true" value of the coefficient will fall outside a range set by its estimated value—2.6 times its standard error. It is highly unlikely, therefore, that the true value of the coefficients of x_1 and x_2 are zero. Or, to put it differently, we can be quite sure that the true value of x_1's coefficient is somewhere between 0.395 and 1.181, and of x_2's, between 0.566 and 1.232

FIGURE 2

ACTUAL COMPANY RATE OF RETURN ON TOTAL CAPITAL COMPARED WITH PREDICTED RATE OF RETURN



return predicted on an investment with no industry dispersion and no fluctuation of company earnings over time. Considering the fact that industrial plant and equipment are not as readily marketable as government bonds, the 5.3-percent level is not an unreasonable estimate of the required return on such an investment.

It is perhaps worth emphasizing that our correlation coefficient, statistically significant as it may be, explains only a small fraction of total variability—specifically only 25 percent. The explanatory power of our equation is illustrated in Fig. 2, which compares the actual return with the rate of return predicted by equation (b). If the relationship was perfect, all points would lie on the diagonal. The scatter of the points indicates the remaining error.

If we turn our attention to equation (e) in Table 3, we note that it bears a fairly close resemblance to equation (b). In fact, the differences between the two are statistically insignificant. That is, given the variability in the data, it would not be surprising to get two equations with coefficients that are similar even if x_2 and x_4 were in fact exactly the same variable. This means that x_2 and x_4 are sufficiently alike that it makes little sense to include both of them in the same regression. This is emphasized in equation (d) in Table 3; when they are included together, the coefficient of x_4 becomes negative and not significant, and the coefficient of x_2 increases. If we consider them both to be one variable, the coefficient in (d) would become 0.861 (i.e., 1.063–0.202), or very close to the values for x_2 and x_4 separately.

Equation (c) was an attempt to measure the effect of skewness of return on the average rate. Not only does it add very little to the correlation, but the sign of the coefficient is the opposite of what we would expect, and its value is just barely statistically significant (less than one chance in one hundred of this result being an accident). We expected to find that companies with returns skewed to the low side would be considered more risky than firms with the same standard deviation but skewed to the high side. When the distribution is skewed to the low side, x_3 is negative, so in order for skewness to increase rate of return its coefficient should be negative as well.

These results may simply mean that there is no relation, but they may also be due in part to the correlation of 0.165 between x_1 and skewness, or the correlation of 0.108 between x_2 and x_3 ,which, as in equation (d), may steal the thunder of x_3. That is, any relation that should be there is already covered in the first two variables. Or it may be a conceptual error—that is, the skewness we were really interested in may be the probability of making losses or rates of return less than the interest rate or some other relevant lower bound. Finally it may be a spurious result due to the fact that companies with negative skewness had low rates of return in some years which pulled down the average rate of return.

Conclusion

■ This finishes our discussion of the main attempts in the direction of a statistical implementation of the Supreme Court's implicit assumption that there is, indeed, a relation between business risk and the rate of return. We are grateful for this opportunity to make more widely available the results of our previously unpublished study. This is not the place, however, to abuse the editors' hospitality by taking the opportunity to answer the various critiques of that study. We leave that for another time.

We cannot stress too strongly the exploratory and tentative nature of this study. The findings are based on a particular body of data, by no means consistent within itself (for example, they embody numerous industry and firm variations in accounting practice, so even rate of return is ambiguously defined), certainly not exhaustive, and probably not fully representative of the universe of industries or firms.

Finally, and perhaps most important, we have considered only business risk. A full study of the relationship between risk and rate of return would require examination of the risk borne by the typical stockholder as well as that specific to the firm. Short of such a full study, however, it may still be useful to examine some of the conditions which might leave an important role for non-financial risk.

Clearly, in a world without organized capital markets one would have to weigh the risks of each capital investment separately. It would be extremely difficult to put together a diversified portfolio without the opportunity to buy small fractions of many large firms through the instruments of bonds and common stock, let alone to objectively estimate the fluctuations in "value" over time of such a portfolio with changing financial and business conditions. Without diversification investors in each business would have to estimate the risks of failure in each endeavor. The line of argument of our paper is simply that in such circumstances the range of outcomes of previously established enterprises is a useful estimate of the risks facing a new enterprise.[4] We argue that, in the absence of opportunities for financial diversification, this measure, so-called "cross-sectional risk," makes much more sense than time-series variability of returns. If a company's earnings displayed highly negative serial correlation of earnings changes alone with a large variance, it would rank as a risky enterprise despite the fact that it was unlikely to face two bad years back to back. There is no economic theory to suggest that earnings changes should not be serially correlated corresponding to the random walk theory of stock prices: indeed, the theory of pure competition suggests such (negative) serial correlation.[5]

Still, once we admit the possibility of diversification through perfect capital markets, there is no question that we can theoretically achieve the industry average rate of return with certainty by buying a fractional share of each entrant in the industry. In such a case, the dispersion of intra-industry returns cannot be thought of as a risk.

No logical relation can establish the empirical validity of an hypothesis. Despite these logically impeccable conclusions, the empirical question remains cloudy. There are regrettable but measurable discrepancies between the capital asset pricing model and the behavior of actual capital markets.[6] Some of these discrepancies are undoubtedly related to either government intervention in the capital markets (e.g., margin regulations) or to transactions costs. Possibly, other discrepancies arise from some species of non-rational behavior. Whatever imperfections do exist leave room for the operation of "risk" factors like those which we attempt to measure.

In particular, we would like to point out two sets of phenomena which might create a non-negligible role for the view of risk discussed in this paper. The theory of financial risk arises from the fact that a perfect capital market permits the elimination of all risk except market risk through diversification. Perfect diversification is achieved by holding a pro-rata share of every asset in the economy. For any individual, to accomplish this might require holding fractional shares of common stock in some firms. Even if we could avoid the fractional share problem, such portfolios would be costly to purchase and supervise, but as an empirical matter, it is well known that we can get very close approximations to the ideal portfolio by holding as few as twenty securities. However, to get this close approximation to the fully diversified portfolio with so few securities, it is critically important that the securities not include more than one from any industry. The greater the industrial concentration of a portfolio, the larger must be the number of different issues held in order to achieve a given approximation to an ideal portfolio. This fact is given recognition in many financial institutions by explicit rules against concentration in industry groups.

---

[4] In practice, we observe only the outcomes of *surviving* enterprises, so that our study makes some implicit assumptions about the relation of failure to survival.

[5] Empirically, earnings changes do seem to behave much as a random walk, which is probably why the "cross-section" and "time-series" approaches do not give such widely varying results, but such luck seems insecure support for a theory.

[6] See, for example, Miller, M. H. and Scholes, M., "Rates of Return in Relation to Risk: A Reexamination of Some Recent Findings" in *Studies in the Theory of Capital Markets*, New York, Praeger (forthcoming).

To the extent that investors find it costly to diversify within, as well as between, industries it will be necessary for industries with especially high intra-industry variability in rate of return to earn premium rates of return to compensate for such costs.

A second obstacle to intra-industry diversification stems from the prevalence of concentrated ownership of securities. For most of its existence, for example, it was impossible to hold a fully diversified equity portfolio of investment in the auto industry because of the family ownership of Ford Motor Company stock. Private issuance of debt is even more widespread and an even greater obstacle to full diversification. We do not wish to make too much of this point: it is possible to substitute highly correlated securities for some monopolized ones, and the opportunities for premium returns in such industries will be limited by inter-industry competition. Still, the obstacle is there. Avoiding it is costly, and costs of this type create opportunities for risk-return relations of the kind we have found.

Whether these considerations are sufficient to establish a more significant relation between return and our measure of risk than that predicted by capital-asset pricing theory or whether the truth lies somewhere between is an unanswered empirical question. This long-interred study of ours is exhumed in the hope it will stimulate further research towards the answer.

Appendix:

Stock price fluctuations and risk

■ As we indicated above, risk arises basically from the unpredictability of the future. Unfortunately, however, it is almost impossible to determine what level of predictability a manager enjoyed. One possible alternative, however, would be to substitute some measure of the predictive ability of another class of knowledgeable participants in the business process—financial investors. While it is impossible to detect in advance which of the earnings fluctuations can or cannot be foreseen, there is a basis for establishing the degree of foresight in any well established financial market. The price quoted for an asset reflects the present value of a future stream of expected earnings. Any unforeseen change in those earnings requires a re-evaluation of the price of an asset. The magnitude and frequency of such changes as occur then become a measure of the frequency and size of the unforeseen changes—in short, a measure of risk.

There are several pitfalls in any study of this kind. First of all, a minimum requirement is that the financial markets to be studied must be reasonably "perfect." What is meant by perfect is that they promptly reflect each piece of new information as it occurs, that there is no sluggish response that would drag the response out over a long period. On this point, we had initially available considerable evidence to suggest that stock markets were quite good performers in this respect, and while we found some contradictory evidence in the course of our research, it seems likely that market imperfections were not a serious obstacle to this approach.

A more serious problem arises if the stock market measures other things in addition to unpredictable earnings fluctuations. Some of the other things are not too important. For example, if investors feel that changes in the country's political complexion alter their judgment about the probability of adverse earnings outcomes, we would like to include such discounting in our measure of risk. On the other hand, we would not like to include risks which accrue solely to the investors and not to the investment itself, such as a threat to the liquidity of the market, a change in capital gains taxation or dividend withholding, etc. We cannot, of course, be sure other factors will not affect our measures and any results must be considered in that light.

Another serious problem arises because the earnings accruing to stockholders are not the same as the earnings we are using to measure rate of return. Our measure includes the return earned by debt holders, and stockholders have only a marginal interest in the interest payments, an interest aroused by the greater variance of earnings induced by debt. The measure we use avoids the risks imposed by financing, while the stock market measure is inextricably related to such risks.

It is theoretically possible to avoid this complication by measuring changes in the market value of a company instead of in its stock alone, but the data gathering difficulties were so substantial that we made no attempt to implement this approach in its pure form.

But we have not neglected it completely. Quite fortunately we were able to get hold of a body of market value data, and we could not resist the temptation of running a few regressions on them.[7]

Specifically, we used equations of the form of the all-years regression in Table 1 and (b) of Table 3, with I and Y defined as before, but with X_{1M} and X_{2M} defined as standard deviations of the rate of return based on market values rather than the book values. X_{1M}, therefore, is the sample standard deviation of investor's rates of return on market value for each company around the unweighted industry average—with return defined as the sum of interest, dividends, and change in market value, and the market value of the company taken to be the market value of its stock plus the par value of its long-term debt. Similarly, market values are used in deriving the standard deviation of the annual rates of return for the company around its postwar average. Note, however, that the rates of return, Y and I, are still based on book value.

---

[7] Every silver lining has a cloud. We note specifically that the three sets of regressions discussed in this section include industry 88—the Miscellaneous group—while those discussed earlier in the body of our paper did not. Strictly speaking, therefore, the findings reported below are not comparable with those in the text Practically, however, we think meaningful comparisons are possible.

The regression equations are:

Industry: $$I_B = \underset{(0.093)}{0.245X_{1M}} + 9.8 \qquad R = 0.383$$

Company:

a) $$Y_B = \underset{(0.064)}{0.263X_{1M}} + 10.1 \qquad R = 0.204$$

b) $$Y_B = \underset{(0.067)}{0.170X_{1M}} + \underset{(0.035)}{0.152X_{2M}} + 7.92 \qquad R = 0.293$$

(In these and the rest of the equations in this appendix we use subscripts B and $\%$ to denote book and market values, respectively.)

In both equations, R is significant and not inconsequential. Yet X_{1M} and X_{2M} are able to "explain" much less of the variation in I and Y than X_1 and X_2. This is why the constant terms tend to be higher than in earlier regressions. This is not surprising in view of the heavy weight that stocks play in the total market value of the firm, and the strong possibility that the price of stock will be determined by more factors than simply earnings prospects. The coefficients of the X's tend to be smaller than before, emphasizing the fact that the variability of market returns is much greater than the variability of book returns. Nevertheless it is worth pointing out that the predictions for company rates of return using X_{1M} and X_{2M} in their regression equation correlate very closely with the predictions of equation (b) in Table 3.[8]

Once we had the data arranged for the computation of the regressions indicated above, it proved to be a simple process to present a number of other relationships as well. For example, while the regressions just discussed evaluate the possibility for predicting rate of return on *book* value given the standard deviations based on *market* value, the following regressions reverse the process.

Industry: $$I_M = \underset{(0.405)}{0.832X_{1B}} + 8.80 \qquad R = 0.309$$

Company:

a) $$Y_M = \underset{(0.210)}{1.056X_{1B}} + 8.15 \qquad R = 0.249$$

b) $$Y_M = \underset{(0.211)}{1.026X_{1B}} + \underset{(0.160)}{0.177X_{2B}} + 7.57 \qquad R = 0.255$$

In these regressions, I_M (industry) and Y_M (company) represent the ratio of dividends plus interest plus changes in market value to beginning market value. The variables, X_{1B} and X_{2B}, are the same ones used in the regressions cited in Tables 1 and 3. It is immediately obvious that the constant terms are of the same order of magnitude as, but a little higher than, the earlier results, the industry equation being most similar to the earlier formulation. Again the correlation coefficients average out lower, and this time the coefficient of X_{1B} in the industry equation is barely significant.

---

[8] That is to say, the correlation for the rankings of companies by each regression is very high. And this tends to confirm our earlier results and strengthen our faith in them. For in the alternative formulation, by taking account of stock price changes and relating a book rate of return to measures of risk not directly related to the dispersion of book rates of return, we have nonetheless come up with a ranking of companies that accords very closely with our previous results.

A final calculation gives some of the most intriguing results. In this last set, we pitted market rates of return against market measures of risk.

$$\begin{array}{lll}
\text{Industry:} & I_M = \underset{(0.480)}{0.832}X_{1M} + 7.50 & R = 0.693 \\
\text{Company:} & & \\
\text{a)} & Y_M = \underset{(0.090)}{0.851}X_{1M} + 7.57 & R = 0.409 \\
\text{b)} & Y_M = \underset{(0.086)}{0.499}X_{1M} + \underset{(0.045)}{0.579}X_{2M} - 0.83 & R = 0.644
\end{array}$$

Clearly, the regressions are relatively impressive from the standpoint of correlation. In general the explanatory power is greater than in any other equation set. On the other hand, the coefficient of X_{1M} in the two company regressions is very volatile, dropping from 0.851 to 0.499 when the second variable is added. Furthermore, the constant term drops from a realistic 7.57 to a negative value, implying that riskless firms would gladly invest at a loss. Just what statistical quirk accounts for this, we could not tell, but the generally good results suggest that some further investigation might prove useful.

The economics of railroading: the beginning of cartelization and regulation

Robert M. Spann and Edward W. Erickson

North Carolina State University

This paper contains an analysis of the need for and effect of regulation in the railroad industry at the inception of the Interstate Commerce Commission. The authors briefly review and integrate a substantial body of literature that has developed over a period of more than half a century. Next they examine the cost conditions under which railroads operated in the latter part of the 19th century and disentangle the separate effects of technological change and economies of large output. Finally, they estimate approximate elasticities of demand in the long and short-haul markets in the trunkline cartel area. These elasticities are used to calculate the relative magnitudes of the gains and losses resulting from the behavior of the railroads in response to the particular form of rate regulation imposed by the ICC.

The classic example of federal regulation in the United States is the experience of the railroad industry under the Interstate Commerce Commission. The transportation industry in general, and the railroad industry in particular, are in a condition of stress. Substantial changes in public policy are almost sure to result. In order to assess what changes are now desirable, it is important to understand what regulation has (or has not) achieved in the past, and at what cost.

Conditions prior to the ICC

■ Prior to the formation of the Interstate Commerce Commission in 1887, railroads, faced with large fixed costs and increasing competition from entry of new roads, attempted to maintain rates through a series of pools and cartel arrangements. At least eight such pools were in operation in 1879 alone.[1] However, with few exceptions they all rapidly failed. One such pool, the Northern Trunk Line Pool, formed to control grain rates between Chicago and the eastern seaboard, has been studied in detail. [9] Although the pool had been characterized as "having reached a high standard of efficiency,"[2] it was basically unstable and went through a series of rate break-

---

Mr. Spann holds the B. S. degree from North Carolina State University (1969), where he is currently working toward the Ph D. degree in economics and statistics. His present research interests include the effects of public policy on industrial performance and the application of economic analysis and models to current social problems

Dr. Erickson holds the B. A degree from Pennsylvania State University (1959) and the Ph.D. degree from Vanderbilt (1968) He taught at Vanderbilt before coming to North Carolina State University, where he is now Associate Professor of Economics. His current interests include the application of the principles of economics to public policy problems and the management of private enterprise

The authors are indebted to Rick Mancke for comments He is in no way responsible for any remaining deficiencies, however.

[1] See [8], p. 8.

[2] See [15], p. 588.

downs and subsequent reorganizations between its formation in 1874 and the beginning of federal regulation in 1887. During this period there were three major disruptions and reorganizations of the agreement and numerous minor disruptions.[3]

The cartel was not completely successful for two reasons. Railroads could not control the entire market due to competition from boats on the Great Lakes. Cartels have inherent instability problems even if they do control the entire market. With regard to water transportation, ". . . the conditions of railway operation were such that water competition was a matter for great concern. Every change in the lake situation was at once reflected in the rail rates, violent dislocations at the opening and closing of navigation in the spring and fall being of special importance."[4] Water transportation, especially in interregional trade, was an effective substitute for rail transportation for six months of the year. [4] Even in 1890, all-rail shipments accounted for only about 75 percent of the grain shipments (but often less for other commodities) from Chicago.[5] The railroads had limited advantages, however, in delivering to Pittsburgh, Philadelphia, Baltimore, and Boston directly and in flexibility in small bulk shipments. Where they could take advantage, they attempted to cartelize.

It is well known that cartels can be unstable in both the short and long run. [16] It is almost always possible for one member of the cartel to profit by breaking the agreement. Since a cartel sets price in excess of marginal cost, any firm can produce an output that is greater than its share of cartelized output and earn greater profits than if it restricted output and joined the cartel. The question is whether these "cheaters'" profits are worth it, because they will initiate price-cutting initiatives from the loyal members of the cartel organization.

Cartel stability in the long run is a strategy problem.[6] Any firm in the cartel faces a demand curve for a secret rate below the cartel rate that is more elastic than the demand curve facing the industry. Hence, in the short run, it can earn profits above its share of cartel profits by quoting a secret rate. However, once the secret rate is discovered by the cartel it may be matched or undercut, in which case the cartel breaks down and the cheating firm (and all other firms) will earn profits that are less than its share of cartel profits. Hence, when the short-run gains from cheating (i.e., profits above the cartel share of profits) exceed the long-run losses from the breakup of the agreement (profits below the level of cartel profits), the cartel is unstable. The parameters that will affect the stability of the cartel are basically the cartel's reaction time to cheating and the elasticities of demand for the market, the cheater, and the cheater in the event of a market breakdown.[7] Although there may be instances where cheating is not profitable, if it generally is cartels are inherently unstable. The behavior of the members of the trunkline cartels indicated fundamental long-run instability.

The history of the trunkline cartels before ICC regulation contains three distinct periods.[8] Each period is closed by a climactic attempt to deal with increasingly aggressive competitive behavior aggravated by changes in the structure of the market.[9] The Pennsylvania and the New York Central both opened service in Chicago in 1869.[10] The first explicit cartel agreement was made in 1874 (the years up to 1874 were characterized by informal discussion of rates between the two big roads). Between 1874 and 1883 there were a number of new entrants into the trunkline railroad market, principally the Baltimore and Ohio roads and an extended Erie Railroad. These years were a period of adaptation of various procedures for setting rates, climaxed by the collapse of the private pool so completely that all negotiations stopped. The year 1884 was the first full year of a period of cartel disintegration and deterioration that did not end until 1886–1887, when the police power of the state created orderly markets through ICC regulation.

---

[3] See [9], pp 13 and 14.

[4] See [14], p. 358.

[5] See [4], p. 39.

[6] See [9], pp. 13–24 and pp. 205–218.

[7] A potential cheater must also consider optimal cartel strategies in the face of others cheating and methods of preventing cheating; however, the conclusions are the same.

The first permanent trunkline organization was formed in Saratoga Springs in 1874. Prior to that time the New York Central and the Pennsylvania had merely quoted the same prices on Chicago to New York shipments, and both roads mostly adhered to them. However, at this time another railroad, the Baltimore and Ohio, completed a line between Chicago and the East Cost.[11] It refused to join any rate agreement at first. This resulted in the first of three major cartel breaks between 1874 and 1879.[12]

In 1876, the trunklines attempted to form new agreements, but the New York Central refused to attend the initial meeting. Ultimately, the New York Central did join. As a result of the Central's ambivalence and subsequent rate breaks by the Pennsylvania and other railroads, this first cartel was never completely stable. The New York Central was accused of cheating in 1877 and appears to have earned both long-run and short-run gains from cheating. Its market share was larger than that imputed to it by an agreement made in March of that year. A rate breakdown during the winter of 1878–79 appears to have been caused by the Pennsylvania. And the road appeared to have gained from cheating.[13]

---

[8] For his analysis, MacAvoy [9] breaks the history of the cartel into two periods, 1871–79 and 1880–87. A tripartite division serves our purpose better.

[9] During the entire period, the primary product carried by the trunklines was grain. As late as 1882, 73 percent of the eastbound tonnage consisted of grain. The majority of this grain originated east of the Mississippi and south of Chicago. The majority of traffic was eastbound See Ripley, William Z., "The Trunkline Systems, a Distance Tariff," *Quarterly Journal of Economics*, Vol 20 (1920), pp. 183–210, reprinted in [15].

[10] See [9], pp. 7—8

[11] The impending entry of the B&O was not the only cause for the Saratoga covenant The New York Central and the Pennsylvania were the only roads with direct service to Chicago from the East Coast, although the Erie had connecting service via the lakes and these two railroads. The simplest possible cartel is one with two members. Even this structure is susceptible to strain Some formalization of the agreement would have been a natural consequence of this strain even without the entry of the B&O. See [2], pp. 312–13 and 358.

[12] MacAvoy [9] determines the timing of the cartel breakup by regressing Chicago-New York grain price differentials on the official rate and comparing the results with those obtained by regressing grain price differentials on average rates posted by the Chicago Board of Trade and the lake-rail rate. Furthermore, MacAvoy examines profits of the individual railroads and compares them to profits that they would have earned had they remained in the cartel Also, he examines market shares of individual railroads to determine whether or not they are significantly above or below trend values. In this instance, the actual cartel break covered the period from the winter of 1874–75 to the winter of 1875–76. During this period, Board of Trade rates—those imputed to particular railroads by grain dealers on the floor of the Board—either explained grain price differentials better than official rates or official rates were being reduced along with the Board of Trade rates. Furthermore, it appears that the Baltimore and Ohio profited from cheating. During the first winter, its profits were below what they would have been had the B&O joined the cartel, but were in excess of cartel profits during the second winter. MacAvoy explained this by noting that during the first winter there may have been no lag in detecting the secret rates, but that by breaking the cartel and then establishing a differential rate to Baltimore, there may have been long-run gains from cheating.

In 1879 a new agreement was formed. Some stability returned to the cartel. However, the measure of stability was short. Official rates were cut during most of 1881 and throughout the years immediately prior to the formation of the ICC: 1884, 1885, and 1886.[14] During this period, roads initiating rate warfare appeared to have gained at the expense of loyal firms. But the gains of cheaters did not outweigh the losses of loyal firms, so in general, railroad profits declined. Several railroads built new lines either into Chicago or from an East Coast port to the Great Lakes during this period. The Grand Trunk Line (a Canadian road) began service between Chicago and New England in 1880. In 1883, the Delaware, Lackawanna and Western Railroad completed a line from New York City to Buffalo and could transport grain from Chicago either by transferring tonnage from other roads or from lake steamers. In the same year the Erie completed its own line into Chicago. In 1884, the New York, West Shore and Buffalo Railway introduced service between Buffalo and New York City by building a line to parallel one of the New York Central's lines. Rail rates declined continuously throughout the several years prior to the formation of the ICC.

The need for regulation

■ The need for regulation in the railroad industry at the time of the formation of the ICC in 1887 rests implicitly on the assumption of decreasing costs in the railroad industry. But costs were in fact non-decreasing. Some of the arguments for decreasing costs are invalid or do not apply to railroads. The railroad industry during the 1870's and 1880's did not behave as an increasing returns industry. Cost data for railroads indicate that while there may have been a range of decreasing costs in railroading, most large railroads operated on a portion of their average cost curve that was non-decreasing.[15]

Perhaps the most common argument is that fixed costs are a large portion of total costs. Hence costs are largely independent of output and per unit costs decline over the entire range of outputs.

---

[13] It should also be noted that during this period an unsuccessful attempt at pooling was made. Market shares were established for each firm, but the agreement broke when the New York Central refused to transfer tonnage to other roads that it had gained through cheating

[14] After a rate war in 1881, the trunkline commission formed an elaborate money pool. Each firm made deposits with the commission. If a road received more than its assigned share of tonnage, payments to deficit roads would either be made directly by the surplus road or out of its deposit with the commission. The agreement was relatively successful until a new entrant, the Delaware and Lackawanna, refused to join the pool at the close of navigation on the Great Lakes in 1883. Rate harmony was never again achieved, as one railroad after another broke the agreements and refused to pay other railroads that had deficit tonnage.

[15] The argument for decreasing costs in the railroad industry can best be summarized in several quotes Shinn states "The rate wars, which have of late years so devastated the finances of railroad companies, are all inaugurated and carried on in interstate traffic. They are detrimental alike to producer, transporter, and consumer, they introduce elements of chance into business which should rest entirely upon Supply and Demand . . . " cited in [8], p. 40, from Shinn, W. P., "The Relations of Railways to the State," *Railway Review*, March 13, 1886, pp 121–122.

Ripley intones "A railroad theoretically presents a clear example of an industry subject to the law of increasing returns—that is to say, an industry in which the cost of operation grows less rapidly than the volume of business done." "Railroads: Rates and Regulation," in [15], p. 71.

Harbeson, in criticizing Kolko's acceptance of the competitive norm, states· "Since railways operate under conditions of decreasing unit costs over a wide range of output, an attempt to enforce price competition would be futile in the long run, and even if feasible would not be in the public interest. It would induce wasteful duplication of investment, prevent the achievement of cost reductions associated with the economies of scale and full utilization of plant and encourage unjustifiable discrimination " Harbeson goes on to admit that if costs are not decreasing regulation is not the answer, but states later ". . . price competition had been, and would be, demonstrably injurious to the long-run interests of shippers as a whole no less than to railroads . . . ." See [5], pp. 232–33.

Railroads have large fixed costs, therefore the railroad industry is a decreasing cost industry. This apparently is the line of reasoning taken by Ripley.[16] Large fixed costs are not a sufficient condition for decreasing long-run costs, and the implications of such an argument are testable. Consider a situation in which short-run average variable costs are constant, yet average total costs decline over the entire range of output due to the existence of some fixed costs. Total unit costs will decline regardless of the ratio of fixed to variable costs, although the slope and position of the ATC curve will be determined by the magnitude of fixed costs and the ratio of fixed to variable costs. Alternatively, average variable costs may initially decline before becoming constant. In neither case are decreasing long-run costs necessarily predicted.

Consider, however, long-run marginal costs as either constant or declining.[17] If long-run marginal costs are constant, we should observe a wide range of firms operating with approximately the same costs and returns. If long-run marginal costs decrease, we should expect to observe existing firms expanding without limit. The data examined in the next two sections clearly indicate that the behavioral pattern followed by the pre-ICC railroad industry and the observed relation between cost and output were only consistent with non-decreasing long-run marginal costs.

The hypothesis of increasing returns and the behavior of railroads

■ What behavioral pattern do we expect to observe in an industry subject to increasing returns? First of all, in the absence of antimerger laws, we should not expect to observe cartels. Rather, it will pay for firms to merge since the resultant larger firm will have lower costs than the sum of the constituent firms in the cartel, and hence profits of all the original entrepreneurs will be larger if the merger occurs than if the industry forms a cartel. However, even though there were no antimerger laws during the cartel years, we did not observe horizontal mergers among large-scale firms. Rather we observed firms forming cartels. The New York Central, Pennsylvania, and Erie systems were all products of previous end-to-end mergers of smaller roads. In the Northern Trunkline Cartel, four of the firms had assets in excess of $100 million (the New York Central, the Pennsylvania, the Baltimore and Ohio, and the Erie),[18] yet none of them merged to enjoy unrealized economies of scale.

There was only one merger during this period. The West Shore Railroad was taken over by the New York Central shortly after it built a line parallel to that of the Central. The West Shore had assets of \$100 million[19] and was built for sale to the New York Central. The merger test is not completely conclusive in itself, but to the extent that the Trunkline Committee tried to maintain rates throughout this entire section of the country, we can tentatively infer that had there been long-run economies of scale, we would have observed mergers among large firms.

---

[16] "Railroads· Rates and Regulation," [15], pp. 44–76.

[17] Since the short-run marginal-cost curve cannot intersect the long-run marginal-cost curve from above, long-run marginal costs are either constant or declining. If short-run marginal costs intersected long-run marginal costs from above, firms would expand up to the optimum size plant and then continue to expand by increasing their variable factors of production until one firm constituted the industry.

[18] The size of the New York Central and the Pennsylvania indicated from balance sheet assets somewhat understates their true size since this figure does not include the assets of lines leased by these railroads, nor does it include railroads controlled by the officers or directors of the Pennsylvania and Central. These railroads appeared to have leased other roads to the extent that inclusion of all leased facilities would almost double each of their respective sizes. The New York Central did not have a direct line to Chicago but operated through a road it controlled, the Lake Shore and Michigan Southern. The Lake Shore and Michigan Southern had assets of slightly under $100 million. The Pennsylvania operated in this area through the Pittsburgh, Fort Wayne and Chicago Railroad, which it controlled. This railroad had assets below $40 million. Also, none of these railroads were exclusively involved in merely the Chicago-East Cost transportation. They also used their assets for local transportation. These arrangements were, in effect, end-to-end mergers. See [13], pp. 203 and 257.

A second implication of a declining-cost hypothesis that can be tested empirically is the following. We should expect to observe rapid expansion of large firms in declining-cost industries. Over-investment is often cited as one of the "evils" of "over-competition" in the railroad industry. The following quotes are samples of past writing on this point:

> In the eastern states, the eighties was a period of speculative 'paralleling' of existing lines of road, in order to dragoon the older lines into purchasing the new ones at exorbitant prices. This was done under the guise of affording satisfaction to the popular outcry for competition for a means of reducing rates. Two notable instances were the building of the West Shore road, paralleling the New York Central, and of the Nickel Plate line which similarly ran for miles within a few rods of the Lake Shore across northern Ohio.[20]

and

> The effect of the pools and the standardized tariffs which they brought into being, when enforced, was to check, in part at least, the ruinous competition between the roads . . . . But they by no means solved in a moment the grievous problems of the American railroad throughout the 'seventies.' At that time, they did not even seek to check another highly disastrous form of competition—over-building, the laying down of new lines oft-times without rhyme or reason and inspired only by silly jealousies or vain ambitions on the part of individual roads.[21]

The growth rates of total assets for the largest roads are compared to the growth rate of total assets for all railroads in the U. S. in Table 1. The growth rates are the annual compound rates. With the exception of one period for the Erie,[22] the rates of growth of the large roads are neither absolutely large nor large relative to the total U. S. industry. Because all costs are variable in the long run and because the specious decreasing-cost argument hinges upon the importance of large capital stocks, we should expect to see total assets for the large roads increasing at a faster than average rate if the overbuilding phenomenon was in fact at work. The data in Table 1 are an indication that it was not. The cartel trunklines were concentrated in the Northeast, however, and the total industry growth rate includes the rapidly expanding western roads. Regional asset data are not available, but ton-mileage data suggest that this bias is not too great. In 1890, roads operating in the area of the trunkline cartel accounted for 50 percent of total U. S. ton-mileage, while all roads operating west of the Mississippi accounted for only 25 percent of total U. S. ton-mileage.

---

[19] See [13], p. 203.

[20] See "Railroads: Rates and Regulations," [15], p. 20.

[21] See [6], pp. 100–101.

[22] The Erie is somewhat of a special case because of its complicated financial affairs. The Erie was in receivership during most of this period. Accounting practices, earnings and assets for the Erie were highly variable in a few years. The Erie was well known as "the Scarlet Woman of Wall Street" according to Charles F. and Henry Adams [1]. Our standard source for the orthodox view states, "Announcement of receivership for the Erie . . . has been a customary feature of our commercial crises for half a century" [14], p. 375. In addition, the Erie showed a low nominal rate of return on assets. The low computed rate of return on assets for the Erie may have been largely a result of problems due to conflict over financial control rather than economic viability. The low rate of return for the Erie should probably not be considered evidence for diseconomies of scale.

TABLE 1

ANNUAL GROWTH RATES OF TOTAL ASSETS FOR SELECTED ROADS

ROAD	TOTAL ASSETS IN 1880 (MILLIONS OF DOLLARS)	PERIOD		
		1875–1880 (PERCENT)	1880–1886 (PERCENT)	1886–1890 (PERCENT)
NASHVILLE, CHATTANOOGA AND ST. LOUIS	15	1.9	2.6	5.2
NORFOLK AND WESTERN[1]	29	N.A.	7.8	21.3
MICHIGAN CENTRAL*	37	1.4	3.1	0.6
DELAWARE AND LACKAWANNA[2]*	40	N.A.	0.3	2.4
LOUISVILLE AND NASHVILLE	47	8.8	12.1	4.0
BALTIMORE AND OHIO*	88	0.0	2.6	5.9
PHILADELPHIA AND READING	128	3.7	4.7	5.1
PENNSYLVANIA*	154	0.0	1.2	1.1
ERIE*	162	9.6	0.0	0.3
NEW YORK CENTRAL, AND LAKE SHORE AND MICHIGAN SOUTHERN*	243	1.7	1.8	1.0
ALL ROADS IN U.S. 1880–1890 7.2 PERCENT				

SOURCE: POOR'S MANUAL OF RAILROADS.

*TRUNKLINE CARTEL MEMBER.

[1] THE NORFOLK AND WESTERN WAS FORMED BY MERGER IN THE LAST HALF OF THE 1870'S.

[2] THE DELAWARE AND LACKAWANNA WAS FORMED BY MERGER IN THE LATTER 1870'S.

[3] DUE TO AN APPARENT CHANGE IN ACCOUNTING METHODS OR A REVALUATION OF ASSETS, ASSET DATA FOR THE NEW YORK CENTRAL IN 1875 IS NOT STRICTLY COMPARABLE TO LATTER YEARS HOWEVER THE DATA WAS USED SO AS TO MAXIMIZE THE GROWTH RATE OF TOTAL ASSETS FROM 1875 TO 1880.

The observed relation between cost and output in railroading depends upon firms' choices with respect to plant size as well as the technical relations between cost and output for a given plant size.[23] Cartelization does not automatically destroy the usefulness of examining the data. During a price war, we may observe firms operating at losses with cartelized overinvestment even if marginal costs are rising in each firm.[24]

Suppose an industry consists of n firms and these n firms form a cartel. Further suppose that all these firms have identical cost curves. If q^* is the output that minimizes average variable costs for any firm and the profit-maximizing output for the cartel is less than nq^*, the cartel will minimize costs for that output by leaving just enough firms (plants) idle so that the producing plants are each producing at the minimum point of their average variable cost curves. At outputs above *nq**, the cartel operates so that marginal costs are equal for all firms. If entry to the cartel is free, new firms will enter until excess profits are eliminated and total output will increase. Existing firms may also increase the number and size of their plants. If *nq** is greater than the cartel output, output will remain the same, new entrants to the cartel will simply increase the optimal number of idle firms, and excess profits will fall due to the fact that fixed costs of the cartel will rise.

---

[23] An argument implicit in the overinvestment *cum* cutthroat-competition description of the railroad industry is that firms in this industry can always raise their scale of operations and reduce marginal variable expenses. Hence, the industry cannot reach a competitive equilibrium and "cutthroat competition" and losses will always be observed. A firm operating at capacity level X can increase its capacity level to aX ($a > 1$) and have lower marginal costs for each level of output Therefore, the industry is unstable and can only exist under regulation or monopoly conditions.

The above argument is false, irrelevant, or both. Shifts in short-run marginal cost schedules such that for given outputs lower marginal costs are associated with plants with larger capacities can be derived from cost functions that have increasing long-run marginal costs We know that short-run marginal cost, under the assumption of constant factor prices, is the ratio of the variable factor's price to its marginal product. Hence, if increases in plant size raise the marginal product of the variable factor, short-run marginal costs are shifted downward as capacity is increased However, all this means is that the variable and the fixed factor are complements in production. It does not imply that long-run marginal costs decrease over the whole range of output or that long-run average costs are not U-shaped.

[24] The following paragraphs are a straightforward adaptation of the analysis found in [12], pp 173–205.

Even in cases where entry is free, however, it is not necessarily instantaneous. Since the purpose of a cartel is to restrict quantity supplied from what it would otherwise be, we would expect to see excess capacity during the process through which a cartel is weakened by entry. If the division of cartelized rents is some function of *pro rata* capital stock, we may simultaneously observe excess capacity, entry of new firms, the expansion of existing firms, and rates of return falling to the normal level. The incentive to cheat remains. But if cheating is widespread, total cartelized rents must fall. Hence, in addition to over-investment we will observe losses in the short run, but these observations do not imply that the industry is subject to a declining long-run marginal cost curve, nor do they imply that the industry cannot achieve a stable competitive equilibrium.

If entry does not unleash centripetal forces which destroy the cartel, and if the cartel operates to maximize joint profits, some plants may be idle or all plants may have some excess capacity. If the binding force of private agreements is not sufficient to maintain the cartel, then either a competitive market may be achieved or the cartel rates may be enforced by public authority. If the latter occurs, there are attendant welfare losses and resource costs due to more plants operating than are necessary to produce the restricted output.

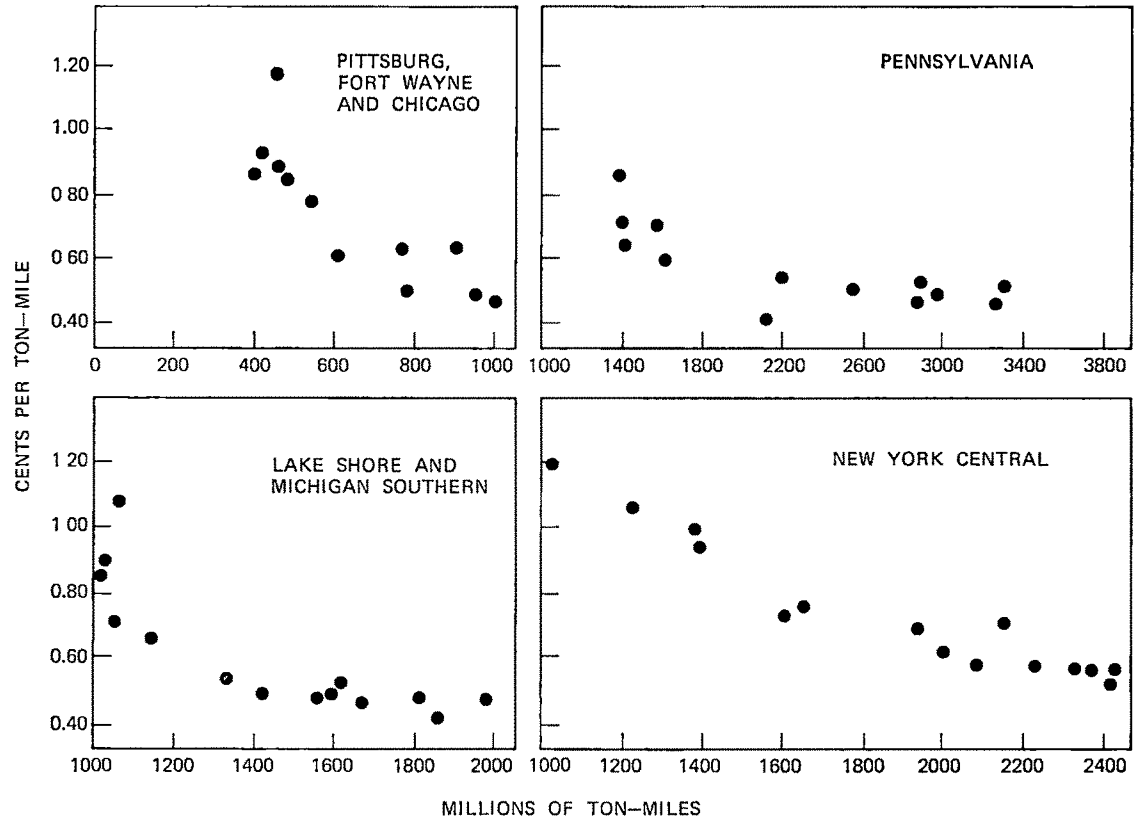
During the period in question, new railroads smaller than existing roads entered the area controlled by the trunkline cartel roads. We should not expect this if the original roads were subject to increasing returns. It is especially noteworthy that the Delaware and Lackawanna, a road of less than half the size of the other roads, expanded and entered the cartel and then broke the cartel agreement. We would not expect a small firm to break a cartel agreement if it had higher costs than the large firms such that once a price war breaks out, the lower-cost firms can undercut the high-cost firm and inflict losses on it without incurring losses themselves. The overall behavior pattern of railroads is not consistent with decreasing costs.

Costs and outputs

■ Examination of cost and output data for the railroad industry, particularly the trunkline cartel firms, indicates that the average cost curve had a declining segment and then became constant. Figure 1 shows average variable costs for four firms plotted against ton-miles of freight traffic for the period 1872–1886. Technical progress was enormous in this period. The effects of technical progress are often mistaken for decreasing costs. We include technical progress in the curves in Figure 1, but show in Figures 2 and 3 that the effect of this is to bias the curves in Figure 1 against the constant cost hypothesis.

FIGURE 1

AVERAGE OPERATING EXPENSES FOR SELECTED ROADS, 1872 THROUGH 1886



SOURCE. POOR'S MANUAL OF RAILROADS

The curves in Figure 1 appear to be the left half of a normal *U*-shaped cost curve. These costs have been corrected for changes in the general price level. Although it appears that railroads were operating on the declining portion of their average cost curves (according to this representation) during this period, the curves seem to be approaching a minimum. The tendency for the plots to show a minimum followed by a range of constant costs is attenuated because some relevant *ceteris* were not *paribus*. These firms did not hold their capital stocks constant. Technical change dominated the results during this period.[25] Hence the curves are biased toward accepting the hypothesis that railroads are a declining-cost industry.

Figures 2 and 3 illustrate the effect of technical change. The ratio of total ton-miles of freight to miles driven by freight trains is plotted against time in Figure 2 and against average cost in Figure 3. Essentially, these plots show that the average ton load per train was increasing during the period and that a decline in average variable cost per ton-mile was associated with this increase. It has been erroneously argued that this implies economies of large-scale operation.[26] This is not the case. Over the long run where the number of trains is a variable, railroads will run that number of trains and tons per trainload which minimizes the costs for any output. In the absence

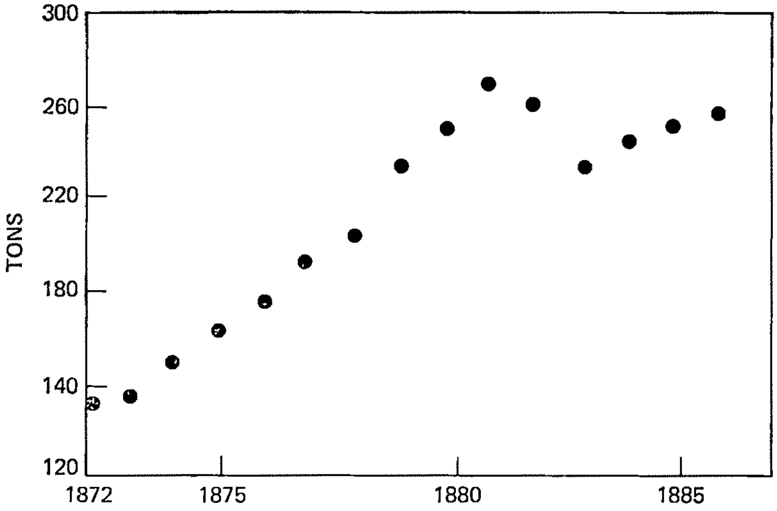
---

[25] See [4], p. 22.

[26] See "Railways Rates and Regulations" [15], pp. 88–98.

FIGURE 2

AVERAGE TONNAGE PER TRAINLOAD FOR THE LAKE SHORE AND MICHIGAN SOUTHERN, 1872 – 1886



SOURCE: POOR'S MANUAL OF RAILROADS

of any improvement in locomotives and rolling stock, tons per trainload would be expected to remain constant. If tons per trainload increase and operating costs per ton-mile decrease over time, this can only be the effect of technological change.

Our interpretation of these data is that the railroad industry was not subject to decreasing costs but was a constant-cost industry in an unstable cartel situation.[27]

The formation of the ICC and its effect on cartel rates

■ The *Act to Regulate Commerce* was passed by the House and Senate on February 4, 1887. Sections 2 through 5 of this act had particularly strong effects on the trunkline cartels. A summary of these sections is as follows.[28]

Section 2: Rebates and personal discrimination of every sort are forbidden.

Section 3: Local discrimination forbidden; equal facilities for interchange of traffic with connecting lines prescribed.

Section 4: Long-haul and short-haul clause; "That it shall be unlawful for any carrier subject to the provisions of this Act to charge or receive any greater compensation in the

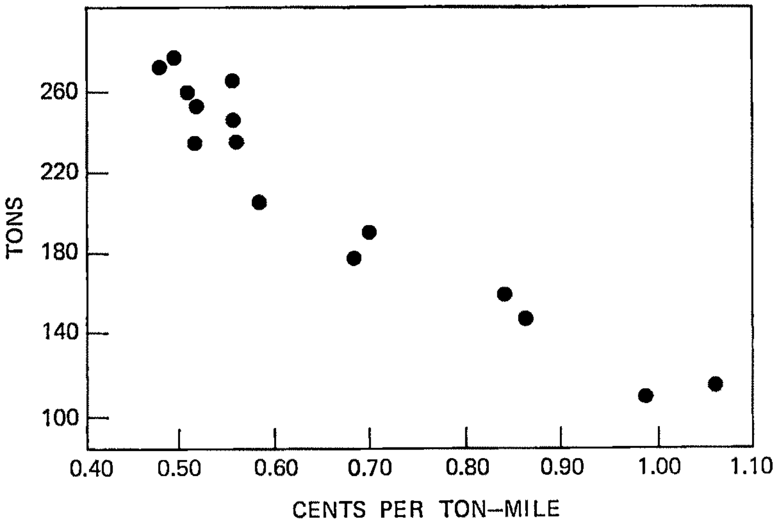
---

[27] Additional evidence to support the constant-cost hypothesis for a wide spectrum of firm sizes was obtained from a sample of 24 firms drawn for the years 1878, 1880, and 1886. Rates of return (rate of return is defined to be total revenue minus total variable cost, all divided by total assets) were plotted against total assets in each of three years. Two of the years (1878 and 1886) followed troughs in the business cycle and coincided with cartel breakdowns Although all the firms in this sample were not in the cartel, three of the larger firms and one of the smaller ones were. The third year, 1880, was a year of cartel strength and a peak in the business cycle. (See [3], p. 28:) However, in all three years the same pattern emerges. Rates of return are low in the small size classes, but virtually the same for medium and large-scale roads. These data are available from E. W. Erickson upon request.

[28] See, for example, "Railroads: Rates and Regulation" [15], pp. 454–455.

FIGURE 3

AVERAGE OPERATING EXPENSES FOR THE LAKE SHORE AND MICHIGAN SOUTHERN, 1872 – 1886



SOURCE: POOR'S MANUAL OF RAILROADS

aggregate for the transportation of passengers or of like kind of property, under substantially similar circumstances and conditions, for a shorter than a longer distance over the same lines, in the same direction, the shorter being included in the longer distance; but this shall not be construed as permitting any common carrier within the terms of this Act to charge and receive as great compensation for a shorter as for a longer distance . . . ." The rest of the section noted that applications for exceptions could be filed with and and would be considered by the Commission.

Section 5: All pooling and traffic agreements are prohibited.

The effects on the trunkline cartel were first to reduce some intermediate rates and selectively raise some long-haul rates so that the structure of rates was in accordance with the law.[29] Short-haul rates fell approximately 15–30 percent, and long-haul rates rose slightly. However, the basic impact of the ICC was to stabilize the cartel and prevent long-haul rates from falling to competitive levels. Since pooling was illegal, new means for cooperation had to be found. The device adopted was the price system. If one railroad was receiving less than its alloted share of traffic, it was permitted to quote a rate differential that would allow it to regain its share. Furthermore, ICC regulations reduced the potential profits from cheating. Any rate cut on long-haul traffic had to be extended to all traffic, and second, any price reduction had to be announced, allowing the other cartel members to meet rate cuts immediately. Up until 1893, although there were some rate breaks, railroads' rates and individual market shares were fairly stable. Even when minor rate breaks occurred, they did not result in complete demoralization of rates as did previous rate wars.[30]

---

[29] See [9], pp. 120–153.

The immediate social costs of ICC regulation

■ During the period that the ICC enforced cartel rates (roughly 1887 to 1893),[31] the economy probably suffered a net welfare loss. The total net welfare loss can be thought of as the sum of a gain from regulation on short-haul traffic and a loss on long-haul traffic. This representation implies a social utility function in which unit gains and losses in each market are valued equally. The reduction in short-haul rates, however, could have been brought about in a manner that did not involve maintenance of cartel rates via the long-haul short-haul clause of the *Act to Regulate Commerce.* An example of such a method is direct price control over short-haul rates. That there were alternatives does not mean that there was one so viable that the costs of achieving the price reduction in short-haul markets was not minimized. Nevertheless, these gains were not costless. Because they are conceptually independent, and because it illustrates the overall effects of one kind of price regulation in interdependent markets, it is worthwhile to compare the relative magnitudes of these gains and losses.[32]

Consider the demands for short-haul and long-haul service. Shipping rates and distances represent a continuum with the per ton-mile rate falling as distance increases (though the total charge, that is the ton-mile rate times distance, rises with distance). Within the range of this continuum there is a distribution of short-haul ton-miles and total charges with a mean shipment and a mean rate. There is a similar distribution for long-haul traffic. We use these mean rates and shipments as representative of the prices and quantities in the long and short-haul markets.

---

[30] There is general agreement with this conclusion. Ripley notes in "Railroads: Rates and Regulation," pp. 454–455, that before the passage of the *Act to Regulate Commerce*, "Published tariffs were only the starting point for the 'higgle' and 'dicker.' It was not bad form for a shipper to 'go shopping' freely among the freight agents of competing lines . . . . By this new statute all was suddenly changed. Rebating was a crime punishable as such . . . ."

Kolko ([8], p. 52) states that ". . . . when the first annual report of the Commission was issued in December 1887, the railroad journalists took the occasion to heap praise on the ICC and the honest men who ran it. At the same time, using the new law as authority the railroads revamped their freight classification, raised rates, eliminated passes and fare reductions, and revised less-than-carload rates on all types of goods, including groceries." Kolko cites a memo of Aldace F. Walker, an Interstate Commerce Commissioner who resigned two years later to take over a railroad traffic association, who wrote ". . . and it is a fact that as a prevention of rate wars and destructive competititon it (the ICC) is already recognized by intelligent railroad men as better than the pool."

MacAvoy ([9], pp. 195–201) and Kolko ([8], p. 57) observe that following the formation of the ICC railroad profits rose during the years 1887–94.

[31] Although there were minor deviations from official rates, there were no serious price wars until 1893. The price wars that began in those years continued through 1895 with only minor interruptions. In 1896 a money pool was formed, but it was short-lived. Concurrently, court decisions were being made which weakened the ICC's power to maintain rates. The most important of these was the Osborne decision in 1892, the Social Circle case in 1893, the Maximum Freight case in 1897, and the Troy case in 1897. The first two cases greatly widened the definition of "essentially different circumstances" in application of the long-haul short-haul clause. The third case questioned the legality of the ICC's actually setting rates and the fourth struck down existing standards for enforcing the long-haul short-haul clause. Furthermore, the Joint Traffic Association, the current name for the northern cartel, was declared in violation of the Sherman Antitrust Act in 1898 [9].

[32] The reduction in short-haul rates did not lower them to marginal costs. The long-haul short-haul clause made the two categories of rates interdependent. The trunkline cartel fixed long-haul rates and the ICC ratified and enforced them. This process also fixed short-haul rates by the intervention of the long-haul short-haul clause. Because market power was greater in short-haul markets, there is no reason to suppose that the profit-maximizing equilibration produced rates equal to marginal cost in the short-haul markets. Comparison of marginal cost estimates with short-haul rates bears out this supposition. Thus, the welfare loss in the long-haul market is an opportunity cost. It is a gain that otherwise would have been achieved in the absence of ICC regulation. The welfare gain in the short-haul market was necessarily contingent upon some kind of regulation, but the actual gain achieved was not necessarily the maximum gain possible. No attempt is made to estimate the possible additional opportunity cost (unrealized potential gain) of inept regulation in short-haul markets. To do so would require that we have an alternative regulatory scheme that (given the skill of regulators and the political attitudes toward income redistribution at the time) would have reduced all prices to long-run marginal costs. We confess that we have no such scheme.

Under conditions of linear demand, the absence of income effects, prices equal to marginal cost in all other industries, and constant costs, the consumer surplus foregone as a result of a price increase is:

$$CS = 1/2 P_0 Q_0 \eta \tau^2 \qquad (1)$$

where P_0 is the equilibrium price, Q_0 is the equilibrium quantity, η is the elasticity of demand, and τ is the relative increase in price.[33]

In the market for long-haul transport, P_0 and Q_0 equal the price and quantity that would have prevailed in the absence of regulation, and τ is the relative change from P_0 to the observed price. In the short-haul market, P_0 and Q_0 are the observed price and quantity in 1890 after prices fell and τ is the relative reduction in price due to the long-haul short-haul clause. Estimates of consumer surplus foregone are for the year 1890. For short-haul transport the ICC regulated price is 2.5 cents per ton-mile.[34] The percentage change in price in short-haul railway transportation due to ICC regulation was about 33 percent.[35] The average rate for agricultural products shipped inter-regionally in 1890 was 0.5 cent per ton-mile.[36] During the 1893 price war, long-haul rates fell approximately 40 percent.[37] Thus we use a τ of 0.67 (τ is a measure of the deviation of market price *above* equilibrium price) and a P_0 of 0.3 cent per ton-mile in the long-haul market under the assumption that price war rates would have prevailed in the absence of ICC regulation.[38]

Quantities were estimated in the following manner. The yearly ICC publication, *Statistics of Railways in the U. S.*, divides the country into ten groups and lists total revenue and average revenue per ton-mile for each road by the group (area) in which that revenue originated. We derived total ton-mileage by road by relevant group from these figures.[39] Since the long-haul short-haul clause only affected shippers that operated in both markets, we shall only estimate the gains and losses on the basis of quantities transported by large lines. Our estimates are proportional to the quantities of traffic handled by the Group II trunkline cartel roads.

---

[33] See, for example, [7], pp. 242–55, and [17], pp 580–94.

[34] This is the average rate calculated by Fogel ([4], p. 72) for intraregional shipments of agricultural products in 1890 It also corresponds to the approximate rate charged for goods shipped in classes 3 and 4 for a distance of 100 miles as published by the Central Traffic Association in the early 1890's and reported in [9], p. 149.

[35] Pre-ICC prices were no more than 33 percent higher than post-ICC prices. The percentage change in price was calculated by taking a small sample of 1886 and 1890 prices for various distances under 200 miles as reported by several roads in the ICC's *Railways in the United States in 1902.* Percentage changes in price were calculated and ranged from zero to over 50 percent. The majority of the observations fell in the 20-30 percent range, so a τ of 0 33 may overstate the gains from ICC regulation.

[36] See [4], p 38. Fogel used the Chicago-New York rates for his calculations This rate corresponds to the 0 25 cents per 100 lbs charged by the cartel on Chicago-New York grain shipments in 1890.

[37] See [9], p 160

[38] It should be noted that we use agricultural commodity rates yet apply our analysis to all commodities. Agricultural commodities accounted for 25 percent of all ton mileage in the eastern U. S. in 1890, although they accounted for a smaller portion of tonnage Grain and other agricultural products, however, accounted for the majority of the trunklines' long-haul shipments They were also generally shipped in lower price classes than other commodities However, to the extent that other commodity rates changed in the same proportion as did agricultural commodity rates, our estimates will not be seriously affected by using these rates.

[39] We calculated the total 1890 ton-mileage for all trunkline roads in Group II (consisting of New York, Pennsylvania, New Jersey, and part of Maryland) and Group III (consisting of Ohio, Indiana, and part of Michigan). These lines accounted for 74 percent and 46 percent of all ton-mileage in Groups II and III respectively We also calculated the ton-mileage for all lines that operated over 400 miles of track in these groups. These roads accounted for 90 percent and 73 percent of all traffic in Groups II and III.

To estimate the ton-mileage of long-haul transporation in Group II, we use the following relationship:

$$S_1R_1 + S_2R_2 = \bar{R}\,, \tag{2}$$

where S_1 equals the share of ton-mileage shipped at long-haul rates, S_1 the share of ton-mileage at short-haul rates, R_1 and R_2 are long-haul and short-haul average revenue, and $\bar{R}$ is overall average revenue. (In all that follows, the subscript 1 always refers to long-haul and the subscript 2 to short-haul.) Letting $\bar{R}$ equal 0.772 (the average revenue for members of the trunkline cartels), R_1 equal 0.5, and R_2 equal 2.5, we obtain S_1 equal to 0.85.[40]

A check is to assume that the distribution of agricultural products between long-haul and short-haul markets is typical of the distribution of all commodities and that all intraregional shipments are short-haul while all interregional shipments are long-haul. Then using Fogel's estimates of ton-mileage of inter and intraregional agricultural shipments as the relevant long and short-haul market quantities, we obtain a second alternative estimate of S_1 equal to 0.73. Because it reduces the losses relative to the gains from ICC regulation, we set S_1 equal to 0.65.

Total shipments over cartel roads in 1890 equaled 18.3 billion ton-miles. Setting S_1 equal to 0.65 and S_2 equal to 0.35 yields estimates of short-haul traffic of 6.4 billion ton-miles and long-haul traffic of 11.9 billion ton-miles. The equilibrium competitive quantity in the long-haul market equals 11.9 billion ton-miles multiplied by $(1 + \tau\eta)$.

To estimate elasticities we make use of the relation

$$MR = MC = P(1 + 1/\eta)\,. \tag{4}$$

We cannot use (4) to estimate directly the elasticity of demand for long-haul service in 1890. This is because the long-haul short-haul clause made 1890 prices in long-haul markets interdependent with prices in short-haul markets. The result of this interdependency is that the observed prices would be on neither the long-haul nor the short-haul demand curve, but on some weighted average composite demand curve. For this reason we use 1880 data to estimate the long-haul demand elasticity.[41] The central estimate of η is 1.5, but this

---

[40] S_1 is not very sensitive to relatively large changes in R_1 or R_2. In addition, we know that

$$\lambda_1D_1 + \lambda_2D_2 = \bar{D}$$

where D_1 and D_2 are the average distances that a ton of commodities was shipped at long and short-haul rates, respectively. The coefficients λ_1 and λ_2 are the proportions of total tonnage shipped at these rates and $\bar{D}$ is the overall average distance a ton of freight was hauled.

We let

D_1 = 500 miles. This is approximately the distance from New York to Buffalo. This figure was was chosen rather than a large distance because this value should represent a mean distance rather than an extreme.

D_2 = 100 miles: Setting D_2 equal to 100 miles makes it slightly less than the average distance that agricultural products were shipped intraregionally as reported by Fogel.

$\bar{D}$ = 190 miles. The average distance a ton of freight was hauled by the New York Central.

Multiplying λ_1 times 500 miles times total tonnage shipped by the New York Central and dividing by total ton-miles yields an alternative estimate of S_1 This alternative estimate is 0.65 which is lower than, but not drastically different from, the first estimate

[41] Cost data are taken from [9], p. 65. Average variable costs are the sum of the costs of locomotive repairs and services, train service, fuel, oil and waste in cents per 100 pounds of grain. Prices are taken from pages 82–83 Marginal costs were computed by MacAvoy on the assumption that they were equal to average variable costs

TABLE 2

SUMMARIZATION OF MARKET PARAMETERS

PARAMETER	SHORT–HAUL MARKET	LONG–HAUL MARKET
P_0	2.5 CENTS PER TON–MILE	0.3 CENTS PER TON–MILE
Q_0	6.4 BILLION TON–MILES	23.6 BILLION TON–MILES
η	1.4	1 5
τ	0.33	0 67

estimate is relatively sensitive to the marginal cost figure chosen.[42] If marginal cost for the high-cost producer is substituted for the majority marginal cost, η becomes 1.7. Because of this, and because the elasticity of demand may have changed during this period, we estimate gains and losses using a range of values for the long-haul price elasticity of demand.

Since marginal costs equal average variable costs, knowledge of overall average variable costs, long-haul marginal costs, and the proportions of traffic carried in the long-haul and short-haul markets will be sufficient to estimate short-haul marginal costs. If $\bar{C}$ is overall average variable cost, C_1 is long-haul marginal cost, C_2 is short-haul marginal cost, and 0.65 and 0.35 are the proportions of total traffic in the long-haul and short-haul markets, short-haul marginal costs are obtained from the following relation:

$$0.65C_1 + 0.35C_2 = \bar{C}. \tag{15}$$

The central estimate of short-haul marginal cost is 0.86 cent per ton-mile,[43] and this yields an estimate of η for short-haul markets of about 1.4. This is in the approximate neighborhood of the estimates of η for long-haul markets and we use the same range of elasticities in estimating gains and losses from regulation.[44] All the central estimates of the market parameters are summarized in Table 2.

Our estimates of the gains and losses resulting from ICC regulation are presented in Table 3. The estimates are for the year 1890 and cover the trunkline cartel area (ICC Groups II and III).[45] The

---

[42] The Chicago-New York grain rate in 1880 was about 30 cents per 100 lbs and marginal cost was about 10 cents per 100 lbs for the majority of the trunklines Marginal cost for the high-cost producer was about 13 cents. We assume that although there was price shading by individual roads the nominal cartel price was set at the profit-maximizing level. Therefore, on the assumption of a constant elasticity at all prices, our estimation procedure yields a true measure of the market price elasticity of demand for long-haul traffic in 1880. The elasticity of demand in 1890, however, may have been different from the elasticity in 1880. This is because the effectiveness of lake steamers as an alternative probably declined and there was substantial population and income growth.

[43] Short-haul marginal costs are expressed in ton-miles rather than in cents per hundred pounds for a well defined route such as Chicago-New York. The average variable costs for the relevant roads ranged from 0 46 to 0 59 cent per ton-mile. This puts limits on $\bar{C}$. Since the low-cost firm has a certain leverage in setting cartel rates, we use 0 46 cent per ton-mile to generate the central estimate of short-haul η. Converting C_1 from cents per hundredweight for Chicago-New York to cents per ton-mile yields a range with bounds set by the low and high-cost railroads. These bounds are 0 20 and 0.25 cent per ton-mile. Together with the extremes of the range for $\bar{C}$, the bounds of C_1 set high and low limits for C_2 equal to 1 31 and 0.86 cents per ton-mile. If the ICC regulated short-haul price was 25 percent below the cartel profit-maximizing price (or the profit-maximizing price 33 percent above the ICC regulated price). The range of the short-haul elasticity of demand is from 1.34 to 1.64

[44] Because there is reason to presume that there were fewer substitutes in short-haul markets—i e , the absence of lake steamers—this suggests that Table 3 should properly be read on the diagonal, from higher right to lower left.

[45] Since equation (1) is linear in ton-miles, a simple extrapolation would appear to give nationwide estimates of gains and losses Because market conditions were not the same in the southern or western regions as in the trunkline cartel area, this cannot be done. The southern market was probably too fragmented for the cartel to be effective, even with ICC enforcement, and the western markets were probably close enough to simple monopoly to make ICC enforcement of cartel rates redundant

TABLE 3

THE GAINS AND LOSSES FROM ICC REGULATION IN THE TRUNKLINE CARTEL AREA, 1890

	TRUNKLINE CARTEL AREA (MILLIONS OF DOLLARS)		
LONG–HAUL LOSSES	13.0	23.4	36.5
SHORT–HAUL GAINS	8.7	12.2	17.4

$\eta_1 = \eta_2 = 1$

$\eta_1 = 1.5$
$\eta_2 = 1.4$

$\eta_1 = \eta_2 = 2$

estimates in Table 3 suffer from the conceptual problems of our technique and the shortcomings of our data,[46] but we feel that they at least capture the relative orders of magnitude of the relevant gains and losses. On this score, ICC regulation was a fiasco. Only if one makes the unlikely assumption that the elasticity of demand in short-haul markets was twice as great as that in long-haul markets do the gains to short-haul customers exceed the losses to long-haul customers.

Implicit in the alternative elasticities used in Table 3 are estimates of alternative quantities demanded at the competitive price for long-haul shipments. For example, if the price elasticity of demand for long-haul transport in the trunkline cartel market was about 1.5 and the observed quantity demanded at the cartelized price was 11.9 billion ton-miles, then a 40-percent reduction in price implies a 7.1-billion ton-mile increase in quantity demanded to 19.0 billion ton-miles. Would the U. S. economy have generated such an additional volume of shipments in response to a price decrease? There is evidence that suggests that the answer is yes.

In 1889, U. S. exports of grain accounted for about 20 percent of shipments of grain in international trade.[47] If the elasticity of supply of grain exports from the rest of the world were zero,[48] then the price elasticity of demand for U. S. exports was about five times the world market price elasticity of demand for grain. If the world market price elasticity of demand for grain was as high as 0.5, then price elasticity of demand for U. S. exports was about 2.5. U. S. grain production in 1887 was 2.6 billion bushels.[49] If all 7.1 billion ton-miles of additional shipments were grain shipments from Chicago to New York, the total additional volume of grain shipped would have been about 230 million bushels. A cut in the freight rate from Chicago to New York would have been the equivalent of about a 10-percent increase in the price of grain at Chicago. The sum of the elasticity of domestic production and the elasticity of substitution for diversion of grain from local to international markets would have had to be about 1.0 in order for grain shipments to provide the whole increase in traffic. Since our assumptions have biased this rough guess of the relevant elasticity upward, and since the whole increase in traffic would not have been grain shipments, this does not seem unreasonable.[50]

---

[46] Some of these shortcomings are, however, overcome by using a range of price elasticities. Both the ratio of long-haul losses to short-haul gains and the absolute size of the difference increases as the elasticity of demand increases. An elasticity of unity is surely the lower limit of the appropriate range (Even the most sanguine exponents of the decreasing-cost hypothesis did not contend that marginal costs were *zero*)

[47] See [11], p 130.

[48] The higher the price elasticity of supply of exports from other countries, the greater the elasticity of demand for U. S exports. We choose a zero elasticity of supply from the other countries to bias the numerical example against the point we are making.

[49] See [11], p. 6.

Conclusions

■ The U. S. railroad industry in the late 19th century was not a decreasing-cost industry and did not require regulation to prevent long-run total real costs from exceeding total revenues. Rather, the observed fluctuations in railroad rates and market shares resulted from the competitive responses of individual firms to the situation of an unstable cartel. The initiation of ICC regulation stabilized the cartel in long-haul traffic and reduced prices for short-haul traffic. It is impossible to tell whether short-haul rates were lowered to rates equivalent to long-run marginal costs, but it is unlikely that the form of regulation imposed achieved this result. The social losses of long-haul customers from cartel stabilization were approximately double the social gains of short-haul customers due to rate reduction. Thus, in the short run, the immediate contribution of the ICC to social welfare was negative.

Regulation often attempts to control the overall performance of a firm or industry by operating on a single key variable. That this can be attempted at all is a result of interconnections between the object variable and the operation of the total system. Ideally, these interconnections do not extend beyond the boundary of that activity which is suited for regulation. [10] The ICC limited monopoly pricing in short-haul markets by linking long-haul and short-haul prices and preventing price discrimination in long-haul markets. In the context of the transport market structure existing at the time, no regulation was necessary in long-haul markets to achieve the results which competitive markets generate. The agent which made the free market effective in long-haul transport was secret price discrimination which rapidly spread to open price competition.

There are two lessons to be learned from this case. Regulation with effects which cut across boundaries between competitive and non-competitive sectors, even if it is successful in achieving its objective in the non-competitive sector, can impose costs in the competitive sector which far outweigh the benefits in the non-competitive sector. Then we are faced with the question whether this is the most efficient and feasible way to achieve the resulting redistribution of income, provided that redistribution is desired. Second, this is another case where the initial agent of competition was secret price discrimination and where legislation to eliminate price discrimination actually solidified a severe misallocation of resources. For obvious reasons, our analysis cannot be extended across time and industries to condemn all such legislation. But imposing anti-competitive controls upon a large market in order to limit monopoly power in a smaller market is seldom likely to result in an increase in net welfare.

---

[50] Fogel, [4] discusses the evolution of the idea of the "Axiom of Indispensability" that railroads were essential to, and the cause of, most of the observed growth in the U. S. economy in the latter half of the 19th century. The references he cites to establish the axiom can be divided into two groups. Those who wrote before the fact and described the potential of railroads, and those who wrote after the fact and made *post hoc, propter hoc* attributions. An example of the former type of reference is J. W. Scott, writing in 1847. "To commercial changes through the interior, it would give an activity beyond anything witnessed heretofore in inland trade. A face of gladness would animate every department of toil, and new motives be held out to activity in enterprise. Social as well as commercial intercourse of the people of distant States would break down local prejudices and annihilate sectional misunderstandings. The wages of labor would be improved, and the profits of capital increased beyond the whole cost of these works . . . . As long as the hills stand, or the valleys disclose their beauty, so long will these works bear evidence to posterity of the energy and spirit of those who erected them."

If our analysis is correct, and taking account of the idiom of the time, Scott was more nearly right than wrong. The alternative explanation of Fogel's *ex post* results is that the cartelized railroads took into account the cross elasticity of demand between their services and the next best substitute and priced accordingly.

References

1. Adams, C. F. and Adams, H. "A Chapter of Erie," *High Finance in the Sixties: Chapters in the Early History of the Erie Railroad.* New Haven: Yale University Press, 1929. Frederick Hicks, ed. Reissued by Kennikat Press, Port Washington, N.Y., 1966.
2. Burgess, G. A. and Kennedy, M. C. *Centennial History of the Pennsylvania Railroad Company, 1846–1946* Philadelphia: The Pennsylvania Railroad Company, 1949.
3. Fels, R. *American Business Cycles 1865–1897.* Chapel Hill, N. C.: The University of North Carolina Press, 1959.
4. Fogel, R. *Railroads and American Growth.* Baltimore, Maryland: Johns Hopkins Press, 1964.
5. Harbeson, R. W. "Railroads and Regulation," *Journal of Economic History,* Vol. 27, No. 2 (June 1967).
6. Hungerford, E. *The Story of the Baltimore and Ohio Railroad.* New York: G. P. Putnam and Sons, 1928.
7. Johnson, P R. "The Social Cost of the Tobacco Program," *Journal of Farm Economics,* Vol. 47, No. 2 (May 1965).
8. Kolko, G. *Railroads and Regulation, 1877-1916.* Princeton, New Jersey: Princeton University Press, 1965.
9. MacAvoy, P. W. *The Economic Effects of Regulation—The Trunkline Railroad Cartels and the Interstate Commerce Commission before 1900.* Cambridge, Massachusetts: The MIT Press, 1965.
10. McKie, J. W. "Regulation and the Free Market: the Problem of Boundaries," *The Bell Journal of Economics and Management Science,* Vol. 1, No. 1 (Spring 1970).
11. Mulhall, M. G. *The Dictionary of Statistics.* London: Routhledge & Sons, Ltd., 1899. Republished by Gale Research Company, Detroit, 1969.
12. Patinkin, D. "Multiplant Firms, Cartels and Imperfect Competition," *Quarterly Journal of Economics,* Vol. 61, No. 2 (February 1947).
13. *Poor's Manual of Railroads, 1887.*
14. Ripley, W. Z. *Railroads: Finance and Organization.* New York: Longmans, Green and Company, 1915.
15. ———. *Railway Problems.* Boston: Grim & Company, 1907.
16. Stigler, G. J. *The Theory of Price.* New York: MacMillan, 1966.
17. Wallace, T. D., "Measures of Social Costs of Agricultural Programs," *Journal of Farm Economics,* Vol. 44, No. 2 (May 1962).

Investment and rate of return for the regulated firm

Blaine E. Davis

Management Sciences Division
American Telephone & Telegraph Company

The subject of this paper is a public utility's optimal dynamic response to the rate of return allowed by a regulatory commission. The problem is formulated in terms of the firm's capital-budgeting policy, and a dynamic analysis is made using modern control theory. This permits the utility's financial policies to be related directly to (a) the way the capital market values the utility and (b) the restriction on earnings imposed by rate-of-return regulation. The solution to the resulting nonlinear control problem provides a quantitative basis for analyzing the impact that small changes in the rate of return have on market valuation and the firm's investment policy. The results furnish one measure of what is commonly called "capital attraction capability," or the firm's ability to attract investment capital in a competitive capital market, and display the relation between that capability and the allowed rate of return.

1. Introduction

■ Because regulatory policy judgements are casuistic in substance they have generally been considered beyond quantitative representation or assessment. However, recent years have seen attempts to apply "scientific method" to the regulatory process through the use of mathematical models. This does not mean human judgement can or should be eliminated. But it may be possible to supplement and buttress judgement so as to improve the fairness and consistency of regulatory decisions.

The purpose of this paper is to present a quantitative framework showing what effect a rate-of-return specification can have on the structure and composition of the regulated firm's long-range investment decisions and the value the capital market assigns to the firm. From this some consequences of regulatory rate specifications can be seen.

The central task is to construct and solve a formulation for a regulated utility in terms of its capital budgeting. This makes it possible to link the utility's investment decisions to (1) the way the capital market values the utility and (2) the restriction that rate-of-return regulation imposes on earnings. This rate-of-return specification appears explicitly in the optimal strategies of the solution to the capital-budgeting model. The allowed rate not only affects the

---

The author is Supervisor—Financial and Regulatory Analysis in the AT&T Company's Management Sciences Division. He received the B. S. degree in Electrical Engineering from the University of Maine at Orono in 1964, and the M S and Ph.D. degrees in Operations Research from Johns Hopkins University in 1965 and 1969, respectively. He joined Bell Telephone Laboratories in 1964 and came to the AT&T Company in 1969. Dr. Davis' current interests are in dynamic models in finance and regulation.

This article is based in part on the author's Ph.D. dissertation. He is indebted to Dr. F. T. Sparrow and Dr. D. J. Elzinga of Johns Hopkins for their advice and direction and to the editors of the BELL JOURNAL for comments and suggestions.

values of the strategies, it also affects the very structure of the solution in that totally different solutions result from different ranges in the rate of return. The impact of small changes in the rate of return on capital budgeting policy and on the market's valuation of the firm has long been observed in practice. This paper furnishes a theoretical framework from which these findings can be analyzed quantitatively.

The results yield one quantitative measure of a "fair" rate of return. It has long been recognized that this rate must be sufficient to attract capital for the growth requirements of utility demand. This appears explicitly in some of the solution cases.

2. Rate-of-return regulation

■ Determining the "fair return" to be allowed a privately owned, regulated company is a complex but important task. In recent times regulatory agencies have regulated earnings by (1) determining the value of the utility's outstanding investment and (2) deciding upon a "fair" rate to apply to that value. The product of value and rate of return is the net revenue constituting a fair return. Price schedules are then specified to yield this net revenue. As could be expected, this vital series of calculations has been subject to an array of political, judicial, economic, and traditional influences whose subtleties almost defy comprehension.

The question what constitutes "fair" in fair rate of return has been the subject of controversy between utilities and government agencies since regulation by statute appeared on the American scene a little over a hundred years ago. Since then a long series of court decisions, culminating with the Hope case, have set forth the "comparable earnings" and "attraction of capital" standards with which the allowed rate of return must comply. An excellent discussion of these standards, and of others implied by them, appears in [17].

The usual way of calculating a fair rate of return is to (1) determine the rate of interest paid on long-term debt, and then (2) combine this with a rate on equity capital. The two must be weighted according to the relative proportions of the two components in the utility's total capitalization. The heart of the rate-of-return question obviously lies in the determination of the rate on equity capital. The complex problem of finding the utility's total capitalization is not treated here. It is assumed that the assets of the utility are properly valued as equal to their opportunity costs in production in some other activity in the economy.

3. General description of problem

■ Attempts to analyze mathematically regulation's impact on the operations and investment policies of utilities can be generally divided into two classes. First, there are formulations using the neoclassical economic models of the firm in terms of production functions and profit maximization. Second, there are formulations concerned with static analysis of the utility's financial structure; models of this type characteristically involve no optimization operations.

Probably the best-known paper in the first class is one by H. Averch and L. Johnson [2], who use the neo-classical model of the firm to demonstrate that rate-of-return regulation tends to produce capital-intensive investment decisions when the firm's objective is profit maximization. Westfield has presented a similar model [24], analyzing the effects of changes in the price of capital and demonstrating conditions under which overinvestment is optimal. Recent investigation by Tasch [23] using this same framework showed the crucial nature of many of the assumptions in [2] and [24].

Among financial models of regulated utilities, a paper by Sparrow [22] presents an analysis specifying what the rate of return would have to be to support specified growth rates of financial indicators recognized as important by investors. Sparrow's model requires no optimization operations and is a generalization and synthesis of two earlier papers [4] and [20].

An econometric approach has been presented by Gordon [7]. His financial model does involve optimization. The formulation is a static analysis on an econometric model of share price.

A recent published attempt to construct a financial model of a firm that integrates capital budgeting with the firm's valuation is a paper by Lerner and Carlton [16]. Their formulation is a static analysis using an investment-opportunities schedule under common-stock price maximization. Their results have been criticized, partly because they depend on the specific form of the investment-opportunities schedule the authors used.

There appear to be no published efforts at a dynamic financial analysis of the firm comprehending both market valuation and capital-expansion considerations. Such an effort, for the special case of a regulated public utility, is the purpose of this paper.

The basic economic problem is to determine the amounts of retained earnings and new equity capital (proceeds from securities issues) allocated to investment under rate-of-return regulation. Investment will increase future earnings. But retention of earnings reduces current dividends, and new equity issues tend to dilute the current owner's equity. Thus, the management's control problem is to choose the investment program that most benefits the owners.

This capital-budgeting problem is formally posed in section 4 as a problem in optimal control. The model is a dynamic, continuous-time model of the utility, encompassing operations and investment in an analysis of financial activity. The formulation is entirely financial in nature, which distinguishes it from an economic model of the firm in that it subsumes the production function and assumes that the utility operates along its optimal expansion path. Optimal expansion in this analysis is concerned exclusively with capital accumulation and its development over time.

The choice of the optimal investment program is constrained by two differential equations describing the change in stock price and equity per share (net worth divided by the number of shares outstanding). These constraints include behavioral assumptions pertaining to market valuation and the utility's operations. Retained earnings and/or new outside capital increase the firm's investment capital. The maximum investment in one period is limited by an upper bound, which can be regarded as the maximum rate at which the utility can efficiently increase its capital and still generate its allowed rate of return.

4. The model

■ Three fundamental differential equations are developed. These show the changes in financial structure under conditions of internal and external financing. The utility chooses the investment program that is most beneficial to the owners, and this is stated in mathematical terms. The development uses the following definitions:

$B(t) =$ total equity,
$N(t) =$ number of shares of common stock,
$E(t) =$ equity per share $(= B(t)/N(t))$,
$X(t) =$ total net earnings,
$d(t) =$ dividend per share,
$r(t) =$ rate of return to equity capital,
$P(t) =$ market price per share of common stock.

All of these variables are functions of time, t.

First, consider the effects of investment on equity. Retained earnings are an effective increase in equity and can be expressed as

$$X(t) - N(t)d(t)$$

or, equivalently, as

$$N(t)[E(t)r(t) - d(t)] .$$

Equity can also be increased by issuing new stock. Let δ denote the discount on market price to enhance the marketability of the new issue. The quantity δ should be interpreted as the average or effective discount, since discounting can take many forms. Internally it represents the cost per share of marketing the issue. Thus

$$0 \leq \delta < 1 ,$$

with

$$(1 - \delta)P(t) = \text{the effective price of the new issue,}$$

and

$$(1 - \delta)P(t)\dot{N}(t) = \text{dollar value of new outside equity raised.}^1$$

Thus, the change in equity can be expressed as

$$\dot{B}(t) = N(t)[E(t)r(t) - d(t)] + (1 - \delta)P(t)N(t) .$$

It follows directly from the definitions that

$$\dot{B}(t) = N(t)\dot{E}(t) + \dot{N}(t)E(t)$$

is an alternative form of this change. Combining these two expressions yields

$$\dot{E}(t) = E(t)r(t) - d(t) + [(1 - \delta)P(t) - E(t)]\frac{\dot{N}(t)}{N(t)}$$

which denotes the change in equity per share. This equation describes the change in owner's equity when the firm both retains some earnings and also raises new equity capital. The only behavioral assumption is embedded in the assumption that δ is a known constant.

Investment by retention and equity financing also affects the price of a share. Unlike the equity equation above, the price equation to be developed is almost exclusively behavioral. It is an attempt to represent mathematically the mechanism of market valuation of publicly owned firms.

---

[1] A dot over any variable denotes the first time derivative of that variable; i.e., $\dot{N} = dN/dt$.

Proceeding along traditional lines,[2] the market mechanism can be stated as follows. The marketability of a stock is governed by (1) the expectation of post-purchase price appreciation and (2) the present level of dividends. This mechanism operates in any trading period to adjust the price of a share to equal the combination of expected price appreciation and current dividends, discounted by a rate applicable to investments of comparable risk. The following traditional assumptions will facilitate a more formal representation of this mechanism.

1. Investors always prefer more wealth to less and are indifferent between dividends and capital gains.

2. No single buyer or seller of shares is large enough to influence the market price significantly. Also, no special costs are incurred in the actual trading mechanism, and there are no tax differentials that could influence preference between dividends and capital gains. (The assumption that any new stock issued by the utility at a discounted market price will not affect the prevailing market price is implied.)

Under these assumptions, and given sufficient time for trading, the following must hold [20]:

$$\hat{P}(t+1) - P(t) + d(t) = \rho(t)P(t)\,,$$

where

$\hat{P}(t+1) =$ the expected price at the end of the trading period (assumed one unit in length)

and

$$\rho(t) = \text{investor discount rate}\,.$$

This is an equilibrium condition and is called the "principle of valuation." It states that in any trading period the market will, given enough time, adjust the price so that dividends plus expected capital gains equal the rate of return the investor requires on an expenditure of $P(t)$.

The "sufficient time for trading" condition means that if

$$\hat{P}(t+1) - P(t) + d(t) - \rho(t)P(t) > 0,$$

then the market will react to increase the current price, and if

$$\hat{P}(t+1) - P(t) + d(t) - \rho(t)P(t) < 0,$$

then the market will react to decrease the current price.

To capture the above character of these equilibrating price adjustments, which counteract tendencies for inequality in the principle of valuation, a continuous-time representation will be used. The following differential equation exhibits this character:

$$\dot{P}(t) = c[d(t) - \rho(t)P(t)],\, c > 0\,,$$

where c can be thought of as a trading activity factor denoting how quickly the market responds to tendencies for inequality in the principle of valuation. This is the mathematical description of price change that will be used in the capital-budgeting model. The traditional assumption of "market certainty" has been made.

The rate $\rho(t)$ is often called the investor rate of return or investor capitalization rate for investments of comparable risk. It should not be confused with $r(t)$, the rate of return on equity capital.

---

[2] An example that uses this traditional point of departure is [20].

Normally, there are three sources of capital funds for investment:

1. retained earnings,
2. new outside equity,
3. debt or senior capital.

It is assumed that the utility has the option to use all three sources.

Let $A(t)$ represent the net assets of the utility at time t.

Then

$$A(t) = B(t) + Q(t),$$

where

$$Q(t) = \text{debt at time } t$$

and

$$B(t) = \text{equity at time } t.$$

Define

$$h(t) = \frac{Q(t)}{B(t)} = \text{debt-equity ratio.}$$

Then

$$A(t) = B(t)[1 + h(t)]$$

and

$$\dot{A}(t) = \dot{B}(t)[1 + h(t)] + B(t)h(t).$$

Recall the development of the relationship for change in equity—

$$B(t) = N(t)[E(t)r(t) - d(t)] + (1 - \delta)P(t)N(t).$$

Substituting this into the relationship for change in assets—

$$\dot{A}(t) = [1 + h(t)]\dot{B}(t) + B(t)\dot{h}(t)$$

yields

$$\dot{A}(t) = [1 + h(t)]N(t)[E(t)r(t) - d(t)] \\ + (1 - \delta)P(t)\dot{N}(t)[1 + h(t)] + B(t)\dot{h}(t).$$

Factoring out $[1 + h(t)]B(t) = A(t)$ and using the identity $B(t) = N(t)E(t)$ yields the relationship

$$\dot{A}(t) = A(t)\left[r(t) - \frac{d(t)}{E(t)} + (1 - \delta)\frac{P(t)}{E(t)}\frac{\dot{N}(t)}{N(t)} + \frac{\dot{h}(t)}{1 + h(t)}\right].$$

This is how the total assets of the utility will increase when any or all of the three sources of capital are used for investment.[3] The rate of change of the utility's asset base is represented by

$$\frac{\dot{A}(t)}{A(t)} = \text{rate of change of net assets,}$$

and this is the quantity which will be bounded under the assumptions to be imposed on conditions for maximum growth rates.

---

[3] The instantaneous rate of change of assets, $\dot{A}(t)/A(t)$, is composed of the three capital sources available to the utility:

1 $r(t) - \frac{d(t)}{Et}$, the contribution from retained earnings,

2 $(1 - \delta)\frac{P(t)}{E(t)}\frac{\dot{N}(t)}{N(t)}$, the contribution from new equity issues,

3 $\frac{\dot{h}(t)}{1 + h(t)}$, the contribution from debt or senior capital.

The capital budgeting activities of a utility must be centrally concerned with preserving the financial integrity of that utility while attempting to meet the investment requirements to satisfy demand for service. The objective of its investment program must be one that is most beneficial to the suppliers of investment funds—old equity owners for internal sources, and new equity owners for external sources. The composition of any investment program should be structured to produce the greatest increase in the worth of their present holdings. This compensation principle is necessary to ensure confidence in the financial integrity of the utility so as to maintain its ability to attract and retain capital. This can be quantitatively stated

$$\operatorname{Max} P(T)\exp[-\rho(T)T] + \int_{t_0}^{T} d(t)\exp\left[-\int \rho(u)du\right]dt,$$

where (t_0, T) represents the planning period of the capital budgeting program with t_0 representing the present time. (The assumption of investor (and owner) indifference between dividends and capital gains is also implied by this objective function.) This will be considered the objective of management in planning the investment program. The planning horizon, T, is quite arbitrary and can be considered infinite if desired. For the case of a bounded terminal price, infinite planning horizon, and constant market capitalization rate, the objective would appear as

$$\operatorname{Max} \int_{t_0}^{\infty} d(t)\exp[-\rho t]dt.$$

Without the maximization operation this function is the classical definition of the price of a share of common stock.[4] Thus as the planning horizon tends to infinity, the objective of management is to choose investments to maximize the price of its stock, an objective function used frequently in static analysis of firm valuation.[5]

Standard and traditional assumptions have been made concerning the mechanism of firm valuation and the objective of the utility. Assumptions will now be made pertaining to the operations of the utility which yield quantitative instruments of control in capital expansion decisions. Special characteristics of a utility which limit some of the variables previously defined will be noted.

The operating assumptions are:

1. Dividends will constitute a proportion of earnings.[6] That is,

$$d(t) = [1 - u_r(t)]E(t)r(t),$$

where

$u_r(t) =$ the retention rate (a management control variable)

and

$$0 \leq u_r(t) \leq 1.$$

The retention rate is defined as the investment accrued from earnings, expressed as a fraction of current earnings.

2. The dollar value of new outside equity subscribed will be deter-

---

[4] The classical reference is [25].

[5] See [8], [9], and [16].

[6] The importance of dividend policy in financial models is not a settled issue. This model tends to emphasize its role. For another view, see [20].

mined as a proportion of current earnings.[7] This assumption allows retained earnings and new outside equity to be determined under the same conceptual structure. The assumption yields

$$u_s(t)N(t)E(t)r(t) = (1 - \delta)P(t)\dot{N}(t),$$

where

$u_s(t)$ = stock financing rate (a management control variable),

and

$$u_s(t) \geq 0.$$

The stock financing rate is defined as the investment accrued from new equity issues expressed as a function of current earnings.

Limitations and restrictions on variables are:

3. The utility has a binding and meaningful rate-of-return constraint in force. Let

r = maximum allowable rate of return to equity capital.

For this constraint to be meaningful its value must exceed the interest rate on debt, and it is always advantageous for the utility to force its actual rate of return to the constraint. That is

$$r(t) = r \text{ for all } t$$

and

$$r > i \text{ where } i = \text{interest on long-term debt.}$$

4. There is presently in the literature no conclusive preference for describing the mechanism underlying changes in the debt-equity ratios of firms. However, once a debt-equity ratio is arrived at there is a tendency to maintain this ratio in future capital-expansion decisions. A constant debt-equity ratio will be assumed in this analysis. Empirical justification for this assumption can be found in [18] and [19].

5. The rate of growth of assets is constrained from above. Due to the special nature of the services and markets of a utility, this bound is assumed known and constant. That is,

$$\frac{\dot{A}(t)}{A(t)} \leq k.$$

6. The conditions of specified maximum growth rate, constant rate of return on equity, constant debt-equity ratio, and a utility type firm collectively infer that $\rho(t)$ should be very stable. If it is assumed that the market rate of interest on long-term debt is constant over the period of interest, then under most generally accepted hypotheses on what influences changes in $\rho(t)$,[8] it can be assumed that $\rho(t)$ will be constant. That is

$$\rho(t) = \rho.$$

7. Assume that the investor discount rate exceeds the growth rate. That is

$$\rho > k.$$

---

[7] This assumption artificially simplifies the difficulty underlying the choice of size of an equity issue. Alternative views of a similar nature appear in [8] and [20].

[8] See [16] for a recent example.

If this relationship does not hold, the "growth paradox" can be encountered, which is not characteristic of firms in the utility sector.

Assumptions 1 and 2 deal with the internal operations of the utility. They are behavioral assumptions which lead to quantitative specification of managerial instruments of control for capital-budgeting decisions. Assumptions 3 through 7 collectively establish the conditions of stable economic stationarity in that the economic conditions encountered in the future will be much the same as they are at present. Thus the framework of the analysis to follow, while dynamic in context, assumes a stationary economic environment.

The capital budgeting problem to be solved can be descriptively stated: "Find the investment rate and composition which will maximize the present value of the owner's holdings when the company is a regulated public utility." Under the assumptions above, this can be mathematically stated as a problem in optimal control,[9] which is

$$\operatorname{Max} P(T)\exp[-\rho T] + \int_{t_0}^{T} [1 - u_r(t)]rE(t)\exp[-\rho t]dt,$$

subject to

$$\dot{P}(t) = c[rE(t)(1 - u_r(t)) - \rho P(t)],$$

$$E(t) = rE(t)\left[u_r(t) + u_s(t)\left(1 - \frac{E(t)}{(1-\delta)P(t)}\right)\right],$$

$$u_r(t) + u_s(t) \leq \frac{k}{r},$$

$$0 \leq u_r(t), \quad u_s(t),$$

and the initial conditions

$$P(t_0) = P_0 \quad \text{and} \quad E(t_0) = E_0 .$$

The terminal condition is given by the fixed planning horizon T.

---

[9] The control problem results directly from substituting assumptions (1) through (7) into the model developed. This can be demonstrated best by considering the differential equation

$$\frac{\dot{A}(t)}{A(t)} = r(t) - \frac{d(t)}{E(t)} + (1-\delta)\frac{P(t)}{E(t)}\frac{\dot{N}(t)}{N(t)} + \frac{\dot{h}(t)}{1+h(t)}$$

Assumption 1) requires $d(t) = [1 - u_r(t)]E(t)r(t)$, thus

$$\frac{\dot{A}(t)}{A(t)} = r(t)u_r(t) + (1-\delta)\frac{P(t)}{E(t)}\frac{\dot{N}(t)}{N(t)} + \frac{\dot{h}(t)}{1+h(t)}.$$

Assumption 2) requires $u_s(t)N(t)E(t)r(t) = (1-\delta)P(t)\dot{N}(t)$, or $\frac{(1-\delta)P(t)N(t)}{E(t)N(t)} = r(t)u_s(t)$. Making this substitution yields

$$\frac{\dot{A}(t)}{A(t)} = r(t)u_r(t) + r(t)u_s(t) + \frac{\dot{h}(t)}{1+h(t)}$$

Assumption 3) requires $r(t) = r$ for all t

Assumption 4) requires $\dot{h}(t) = 0$ for all t, thus

$$\frac{\dot{A}(t)}{A(t)} = ru_r(t) + ru_s(t)$$

Assumption 5) requires this quantity to be bounded by k, i.e.,

$$\frac{\dot{A}(t)}{A(t)} \leq k .$$

It follows directly from this that the controls $u_s(t)$ and $u_r(t)$ are constrained by the relation

$$u_r(t) + u_s(t) \leq \frac{k}{r}.$$

5. Preliminaries

■ Applications of control theoretic techniques to models of management control systems have not been plentiful. One reason may be the difficulty associated with the synthesis after the candidate solutions have been found. We deal with this difficulty here by an application of the construction technique of [11], cast in the control theoretic structure of the Maximum Principle [21]. This reverse time construction technique is particularly effective in solution synthesis in that the complete solution is "built up" quantitatively by construction from solution cases. The solution cases are identified by treating the analysis by the Maximum Principle as a mathematical programming problem at each instant of time. This integration of the analysis and the synthesis provides the systematic method for solution.

In constructing the solution, the terminal manifold is parameterized in terms of the state variables. By starting here and moving backwards in time, the entire state space is filled with optimal trajectories solving the problem for arbitrary initial states. This reverse time synthesis is guided by examining the solution cases encountered in the application of the Maximum Principle. For this model the solution cases are determined by a linear program solved by inspection. The approach here is similar to that of [6]. See also [11]. The Maximum Principle is also modified by an approach [1] to allow ease in handling discounted objective functionals. To demonstrate this, and to briefly review the general approach, the necessary conditions for optimal control are reviewed below.

Without the exponential discount term in the objective function, the capital budgeting problem would be of the form

$$\max_{U(t)} V(X,t_0) = G[X(T),T] + \int_{t_0}^{T} L[X(t),U(t),t]dt$$

$$\dot{X}(t) = f[X(t),U(t),t]$$

$$U(t) \in \Omega$$

$$X(t_0) = X_0$$

$$T \text{ fixed}.$$

The optimal control function, $U^*(t)$, can be found by applying the Maximum Principle [10], [21], formally stated as follows .Let $U^*(t)$ be the optimal control function which maximizes $V(X,t_0)$. Then for $U^*(t)$ to be optimal it must necessarily satisfy

$$\dot{X}^*(t) = \frac{\partial H}{\partial \lambda},$$

$$\dot{\lambda}^*(t) = -\frac{\partial H}{\partial X},$$

$$\lambda^*(T) = \frac{\partial}{\partial X}[G[X^*(T),T]],$$

$$H = \frac{\partial H}{\partial t},$$

$$X^*(t_0) = X_0,$$

and

$$H[X^*(t), U^*(t), \lambda^*(t), t] \geq H[X^*(t), U(t), \lambda^*(t), t]$$

for all

$$U(t) \in \Omega$$

(vector notation is used), where H is known as the Hamiltonian function defined as

$$H = G[X(t), U(t), t] + \lambda(t) \cdot f\,[X(t), U(t), t]$$

and $\lambda(t)$ is commonly referred to as the adjoint vector. In using the Maximum Principle, one attempts to derive a relationship between the optimal control and the state and adjoint variables. In general this relationship may be expected to yield

$$U^*(t) = g[X^*(t), \lambda^*(t)]\,.$$

If this relationship can be uniquely determined, then this represents the nonsingular case. If, however, the direct application of the Maximum Principle does not lead to a unique determination of $U^*(t)$, then a singular situation results. The most prevalent case is when the Hamiltonian is trivially satisfied for any $U(t)$ over some finite time interval. This is traditionally known as the problem of singular controls. This condition has been investigated in a number of articles, of which [12] is representative. Probably the most extensive treatment of singularities is contained in [11].

When a singular control situation is encountered, additional analysis is required to generate the singular trajectory. It is necessary that

$$\frac{\partial}{\partial U}\frac{d^2}{dt^2}\left[\frac{\partial H}{\partial U}\right] \geq 0\,.$$

This relation is known as the Generalized Legendre-Clebsch Condition: [12] and [14].

The capital budgeting problem as stated, however, has an exponential discount term in the objective functional. The introduction of this exponential into the objective functional can be troublesome in terms of analysis by the Maximum Principle. However, a technique discussed in [1] mitigates this situation and will be treated below.

Recall the definition of $V(X, t_0)$. Let $V^*(X, t_0)$ be the maximum value of $V(X, t_0)$. Then from the Hamiltonian-Jacobi partial differential equation [3] for the optimal V, the following holds,

$$\frac{\partial V^*}{\partial X} = \lambda^*\,,$$

if a unique $V^*(X, t_0)$ solution of sufficient smoothness exists. Now consider the case where the objective function is discounted. That is, consider

$$V(X, t_0) = G[X(T), T]\exp[-\rho T] + \int_{t_0}^{T} L[X(t), U(t), t]\exp[-\rho t]dt.$$

This formulation requires that the returns L and G be discounted to time 0. It is more natural to discount them to t_0 since this is the beginning of the range of interest. To do this, define the current value function

$$W(X, t_0) = \frac{V(X, t_0)}{\exp[-\rho t_0]}.$$

Assuming a unique optimal solution of sufficient smoothness exists, then it appears more natural to define the adjoint variables in terms of the current-value function:

$$\lambda = \frac{\partial W}{\partial X} = \frac{\partial V}{\partial X}\left[\frac{1}{\exp[-\rho t]}\right].$$

Now application of the Maximum Principle to the discounted objective functional yields

$$\hat{H} = \exp[-\rho t]H(X, U, \lambda, t)$$

where H is the Hamiltonian for the undiscounted case. Application of the necessary conditions requires a change only in the adjoint equation. The condition now must be stated as

$$\frac{d}{dt}[\exp(-\rho t)\lambda(t)] = -\frac{\partial}{\partial X}\hat{H},$$

from which the following results:

$$\dot{\lambda}^*(t) = -\frac{\partial H}{\partial X} + \rho\lambda^*.$$

Now equation $\hat{H}$ is maximized by choice of $U(t)$, the control vector, and that is equivalent to maximizing H, the undiscounted Hamiltonian. The introduction of a discount term in the objective functional can be handled by a change in the equation defining the adjoint variable and then proceeding as required in the undiscounted case.

6. Solution of the control problem

■ In order to center attention on the economic evaluation of the solution of the capital-budgeting problem, this section bypasses the analytical complexities of solution synthesis in nonlinear optimal control problems. An outline of the synthesis technique which leads to the solution is presented in the appendix, and the synthesis is available in full detail in [5].

The Maximum Principle provides a means of finding the optimal control function under the implicit assumption that optimal control exists. This question of existence arises since the Maximum Principle requires the existence of the adjoint vector. If optimal control exists, then the existence of the adjoint vector is assured. This adjoint vector has a central role in the actual determination of the optimal control function. In general, existence of optimal control functions for nonlinear systems is difficult to assure. However, the existence of the optimal control function for this problem is established.[10]

---

[10] See [15], page 262, for proof of this statement.

Application of the Maximum Principle with the modification to remove the discount term requires that the optimal u must satisfy[11]

$$\operatorname{Max} H = Er + c\lambda_P[Er - \rho P] + Er[\lambda_E - c\lambda_P - 1]u_r + Er\lambda_E\left[1 - \frac{E}{(1-\delta)P}\right]u_s$$

where

$$\dot{\lambda}_P = (c+1)\rho\lambda_P - \left(\frac{E}{P}\right)\frac{\lambda_E Er}{(1-\delta)P}u_s, \tag{1}$$

$$\dot{\lambda}_E = -r(c\lambda_P + 1) + \rho\lambda_E - r(\lambda_E - c\lambda_P - 1)u_r - r\lambda_E\left[1 - \frac{E}{(1-\delta)P}\right]u_s + \frac{\lambda_E Eru_s}{(1-\delta)P}, \tag{2}$$

$$\dot{P} = c[Er(1-u_r) - \rho P], \tag{3}$$

$$\dot{E} = Er\left[u_r + u_s\left(1 - \frac{E}{(1-\delta)P}\right)\right], \tag{4}$$

and

$$u_r + u_s \leq \frac{k}{r}, \quad u_r \geq 0, \quad u_s \geq 0,$$

with boundary conditions

$$P(t_0) = P_0, \quad E(t_0) = E_0, \quad \lambda_P(T) = 1, \quad \lambda_E(T) = 0.$$

To facilitate the analysis, H may be written as

$$H(\lambda, P, E, u) = h(\lambda, P, E) + S_r(\lambda, P, E)u_r + S_s(\lambda, P, E)u_s,$$

where

$$h(\lambda, P, E) = rE + c\lambda_P(rE - \rho P),$$

$$S_r(\lambda, P, E) = rE[\lambda_E - c\lambda_P - 1],$$

and

$$S_s(\lambda, P, E) = rE\lambda_E\left[1 - \frac{E}{(1-\delta)P}\right].$$

The maximization of H with respect to u can be characterized by the switching functions, S_r and S_s, which are the components of the gradient of H with respect to u_r and u_s respectively.

The feasible controls lie in the triangle bounded by

$$u_r = 0, \quad u_s = 0 \quad \text{and} \quad u_r + u_s = \frac{k}{r}.$$

Thus the maximization of H with respect to u (at a particular time) is a linear programming problem which may be solved by inspection. Depending on the values of S_r and S_s, the unique solutions are on the vertices (the bang-bang solutions) and the non-unique solutions are on the faces or in the interior (the singular solutions). The solution cases are listed below.[12]

---

[11] Throughout the analysis the time argument is omitted unless obvious ambiguity results.

		CONDITION
a)	$S_r > 0\,,\quad S_r > S_s$	$u_r^* = \frac{k}{r}\,,\quad u_s^* = 0$
b)	$S_s > 0\,,\quad S_s > S_r$	$u_r^* = 0\,,\quad u_s^* = \frac{k}{r}$
c)	$S_r < 0\,,\quad S_s < 0$	$u_r^* = 0\,,\quad u_s^* = 0$
d)	$S_r = 0\,,\quad S_s < 0$	$0 \le u_r^* \le \frac{k}{r}\,,\quad u_s^* = 0$
e)	$S_r < 0\,,\quad S_s = 0$	$u_r^* = 0\,,\quad 0 \le u_s^* \le \frac{k}{r}$
f)	$S_r = S_s > 0$	$u_r^* \ge 0\,,\quad u_s^* \ge 0\,,\quad u_r^* + u_s^* = \frac{k}{r}$
g)	$S_r = 0\,,\quad S_s = 0$	$u_r^* \ge 0\,,\quad u_s^* \ge 0\,,\quad u_r^* + u_s^* \le \frac{k}{r}$

The singular solutions (conditions d through g) require further analysis for a unique determination of the control variables (see appendix).

By using the conditions on the switching function, solution synthesis can be performed by the reverse-time construction technique. The solution will be presented by demonstrating the effects on the state variables under conditions of optimal control. Displaying the state variables in terms of price-equity ratios permits a more comprehensive understanding of movement in the state space. The solution will be displayed in a plot of P/E versus time.[13]

In performing the solution synthesis it was found that the solution depended critically on the relationship between r and ρ. Five solution cases will be displayed in terms of the ratio r/ρ. These are

1. $\frac{r}{\rho} \le 1$
2. $1 < \frac{r}{\rho} < \frac{1}{1-\delta}\,,\quad \delta > 0$
3. $\frac{r}{\rho} = \frac{1}{1-\delta}$

---

[12] Cases *a*, *b*, and *c* are the nonsingular solutions, while cases *d* through *g* are singular.

[13] Complete representation of the state space requires P/E, time, and either P or E itself. However, the salient features of the solution can be shown in the P/E versus time plot, which avoids the complexities of three-dimensional plots. These are, of course, price-earnings ratios adjusted by a constant.

4. $$\frac{1}{1-\delta} < \frac{r}{\rho} \le \frac{c + [c(1-\delta)+1]\frac{k}{\rho}}{c(1-\delta)}$$

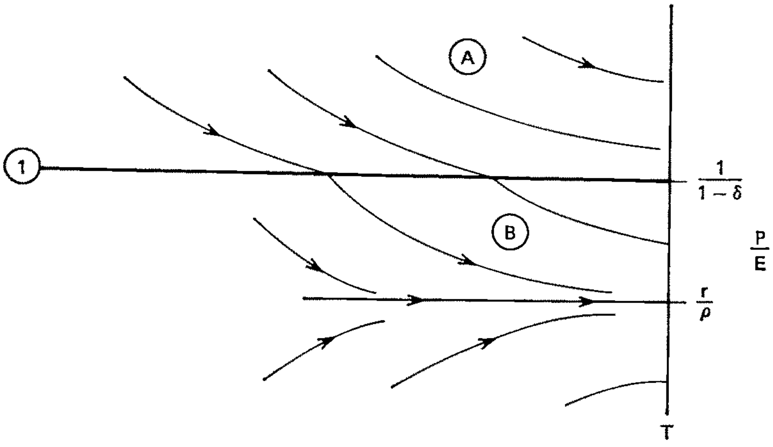
5. $$\frac{r}{\rho} \ge \frac{c + [c(1-\delta)+1]\frac{k}{\rho}}{c(1-\delta)}$$

Each of these cases displays a significant change in the structure of the solution. The vector field representation of the price-equity ratio will be concentrated around the value $1/1 - \delta$. This is where the major changes in movement occur. For ratios significantly outside this range it will be obvious from the vector field how it could be smoothly extended to attain an excluded starting position.

Each solution case will be displayed as a planar vector field with time as the abscissa and price-equity ratio as the ordinate. The terminal manifold, T, bounds the field with starting positions left arbitrary. Capsule descriptions accompany each solution case. A more general discussion of the solution cases appears at the end of this section. Economic interpretation of the solution cases appears in section 7.

The following are the solution cases.

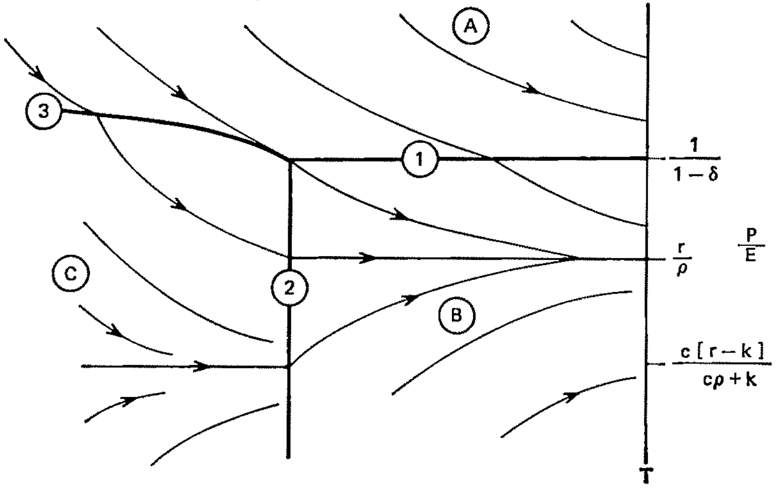
CASE 1) $\frac{r}{\rho} \le 1$



Ⓐ $u_r = 0,\ u_s = \frac{k}{r}$　Ⓑ $u_r = 0,\ u_s = 0$　① $S_r < 0,\ S_s = 0$

Solution case 1 has simple structure with a switching manifold at $P/E = 1/1 - \delta$. All optimal trajectories tend towards the line $P/E = r/\rho$. The trajectories in Region B are asymptotic to r/ρ; they do not intersect it. Thus the price-equity ratio tends towards r/ρ for all t. For sufficiently large T, this position will be approached yielding $\dot{P} = \dot{E} = 0$. There will be no growth, dividends will be equal to earnings and constant, and the equilibrium P/E ratio will be less than or equal to one.

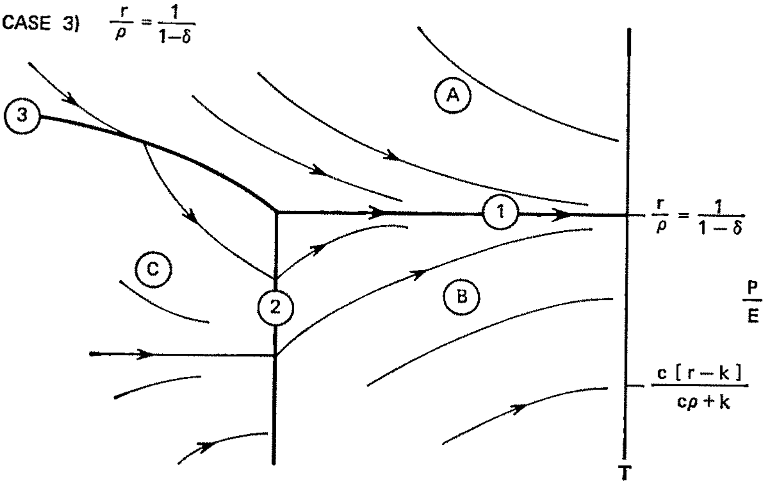
CASE 2) $1 < \frac{r}{\rho} < \frac{1}{1-\delta}$



Ⓐ $u_r = 0, \; u_s = \frac{k}{r}$ ① $S_r < 0, \; S_s = 0$
Ⓑ $u_r = 0, \; u_s = 0$ ② $S_r = 0, \; S_s < 0$
Ⓒ $u_r = \frac{k}{r}, \; u_s = 0$ ③ $S_r = S_s > 0$

The structure of solution case 2 is more complex due to the switching manifolds at $S_r = 0$, $S_s < 0$ and $S_r = S_s > 0$. It should be noted that $S_r = 0$, $S_s < 0$ will be well removed from T. For starting positions in region C retention of earnings is optimal. The optimal trajectories are again asymptotic trajectories in regions C and B. Noting that $S_r = 0$, $S_s < 0$ is well removed from T, this case is essentially the same as case 1 with essentially no growth for the bulk of the capital-budgeting program duration. Observe that in this case the manifold $S_r = S_s > 0$ appears and is a switching manifold.

CASE 3) $\frac{r}{\rho} = \frac{1}{1-\delta}$



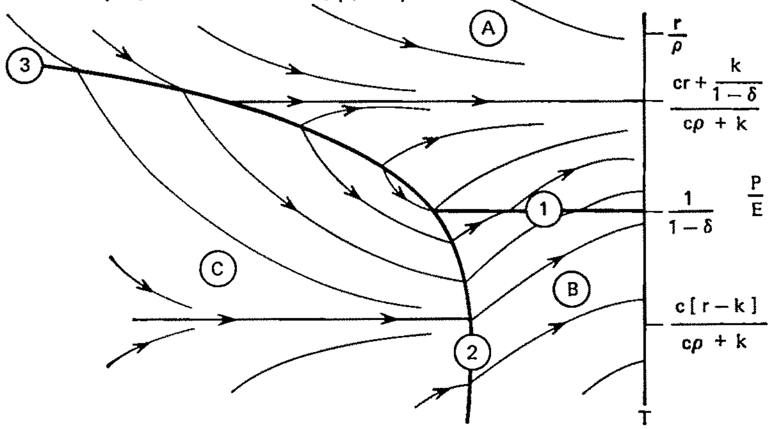
Ⓐ $u_r = 0, \quad u_s = \frac{k}{r}$ ① $S_r < 0, \quad S_s = 0$

Ⓑ $u_r = 0, \quad u_s = 0$ ② $S_r = 0, \quad S_s < 0$

Ⓒ $u_r = \frac{k}{r} \quad u_s = 0$ ③ $S_r = S_s > 0$

Solution case 3 is essentially the same as case 2 except that the manifold $S_r < 0$, $S_s = 0$ is singular and movement on it is optimal. The trajectories surrounding it are asymptotic to it; they do not intersect it. Again the optimal trajectories are asymptotic in region C, as was true in the other cases.

CASE 4) $\frac{1}{1-\delta} < \frac{r}{\rho} \leq \frac{c + [c(1-\delta)+1]\frac{k}{\rho}}{c(1-\delta)}$


Ⓐ
③
$\frac{r}{\rho}$
$\frac{cr + \frac{k}{1-\delta}}{c\rho + k}$
①
$\frac{1}{1-\delta}$
$\frac{P}{E}$
Ⓒ
Ⓑ
$\frac{c[r-k]}{c\rho + k}$
②
T

Ⓐ $u_r = 0$, $u_s = \frac{k}{r}$ ① $S_r < 0$, $S_s = 0$

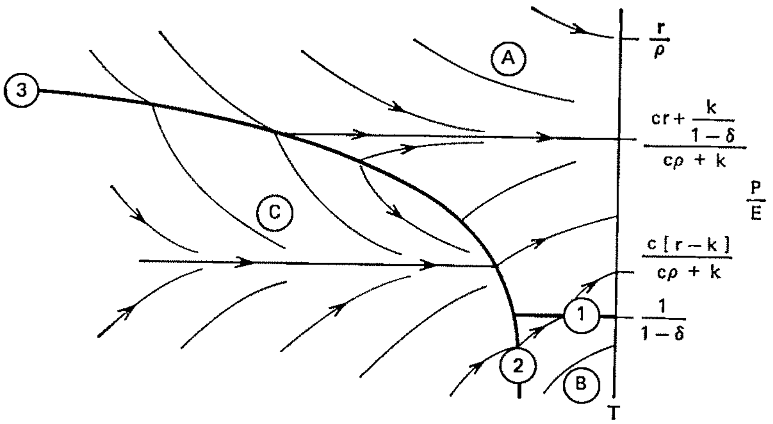
Ⓑ $u_r = 0$, $u_s = 0$ ② $S_r = 0$, $S_s < 0$

Ⓒ $u_r = \frac{k}{r}$, $u_s = 0$ ③ $S_r = S_s > 0$

Solution case 4 shows a marked departure from the structures of cases 1, 2, and 3. In this solution, the manifold $S_r = S_s > 0$ reveals a choice in optimal strategy for a range of starting positions on it. Either choice yields the same value in terms of the objective. Movement along the surface is nonoptimal. Surfaces of this type were first demonstrated by Isaacs [11] in the context of differential games. He has named them dispersal surfaces. The optimal vector field of trajectories in this case is more complicated and clearly demonstrates the nonuniqueness of the solution for this locus of possible starting positions. Although the price-equity ratio decreases along some of the optimal trajectories, leaving the manifold $S_r = S_s > 0$, one should not automatically conclude that the price is decreasing throughout the trajectory length in region C. It can be either increasing or decreasing, depending on the starting position and the parameter values. If it is increasing, it will do so at a slower rate than the equity increase. As before, the optimal trajectories are asymptotic.

Case 5 is a direct extension of case 4 for larger values of r. The essential difference is that now the terminal price-equity ratio will normally be greater than $1/1 - \delta$ for most programs except for very low price-equity starting positions. This case demonstrates the range of r needed to guarantee this condition. Although not a part of this formulation, it is well recognized that price-equity ratios less than one are avoided in practice if possible.

CASE 5) $\frac{r}{\rho} \geq \frac{c + [c(1-\delta) + 1]\frac{k}{\rho}}{c(1-\delta)}$



A
3
C
1
2
B
T
$\frac{r}{\rho}$
$\frac{cr + \frac{k}{1-\delta}}{c\rho + k}$
$\frac{P}{E}$
$\frac{c[r-k]}{c\rho + k}$
$\frac{1}{1-\delta}$

Ⓐ $u_r = 0, \quad u_s = \frac{k}{r}$ ① $S_r < 0, \quad S_s = 0$

Ⓑ $u_r = 0, \quad u_s = 0$ ② $S_r = 0, \quad S_s < 0$

Ⓒ $u_r = \frac{k}{r}, \quad u_s = 0$ ③ $S_r = S_s > 0$

The solution structures can generally be divided into two similar classes, cases 1, 2, 3, and cases 4 and 5. For rates of return in the range of $r/\rho = 1$ (cases 1, 2, 3), the solution structure is quite simple from a control theoretic point of view. The economic significance of rates of return in this range is discussed in the next section. When the rate of return is somewhat higher than $r/\rho = 1$ (cases 4 and 5), the solution structure becomes complex, especially in the region containing the manifold $S_r = S_s > 0$. An economic interpretation of this manifold is given in the next section.

7. Economic interpretation of results

■ Economic interpretation in control applications has been particularly rich in macro growth models.[14] This is generally accomplished by viewing the adjoint variables as time varying Lagrange multipliers. The results of this analysis can be interpreted in an analogous manner by identifying the adjoint variables in a similar vein.

From the investors' point of view the "product" of the utility is the value accrued from share ownership. Let $V(P, E)$ be the "product," where

$$V(P, E) = P(T)\exp[-\rho T] + \int_{t_0}^{T} [1 - u_r(t)]E(t)r\exp[-\rho t]dt\,.$$

Now

$$\frac{\partial V}{\partial P} = \text{marginal "product" with respect to price}\,,$$

and

$$\frac{\partial V}{\partial E} = \text{marginal "product" with respect to equity}\,.$$

From the Hamiltonian-Jacobi partial differential equation for the optimal V (see Section 5),

---

[14] See [1] and [6] as recent examples.

$$\frac{\partial V^*}{\partial P} = \lambda_P$$

and

$$\frac{\partial V^*}{\partial E} = \lambda_E .$$

Thus the marginal products at the optimal V are equal to the adjoint variables. The state variables P and E should here be considered as inputs to a structure which produces the "product," $V(P,E)$. It has been assumed that these inputs are perfectly competitive (the investor is indifferent between capital gains and dividends, with the latter accruing directly from equity). Using the economic argument that in the case of competitive inputs the proper assigned price of an input must be equal to its marginal product, the adjoint variables can be thought of as shadow prices. This gives an economic meaning to maximizing the Hamiltonian at every instant of time as a necessary condition for an optimal solution over time. By their actions via their investment policies, "management" can "give" to the shareholder at each instant of time, dividends, a change in current share price which will effect the terminal price, and a change in equity from which future earnings can accrue. The value of this at time t is simply the sum of the three weighted by their respective shadow prices or values, i.e.,

$$d(t) + \lambda_P(t)\dot{P}(t) + \lambda_E(t)\dot{E}(T) .$$

This is the implicit current value of share ownership, assuming the shadow prices are given. This is also the Hamiltonian of the control problem (see section 6). Thus by maximizing the implicit value of share ownership at every instant of time, the actual value of share ownership over time is optimal. By rearranging the above into the form used in section 6, it is now seen that the switching functions, $S_i(P, E, \lambda_P, \lambda_E)$, provide the utility management with the necessary comparison to make this maximization since their instruments of control are implicit in all the terms. The most interesting case is

$$S_r = S_s > 0 .$$

Here any comparison is inconclusive on how to maximize the implicit instantaneous value of ownership. To resolve this case the "managers" should look at the first time derivative of S_r and S_s. If they are not equal, then the condition is a transient one and in the control theoretic sense the manifold is a switching surface. This condition is encountered in solution cases 1, 2, and 3. If, however, the first time derivatives of S_r and S_s are equal, this test does not reveal the optimal strategy. It is here that the dispersal surface demonstrated in solution cases 4 and 5 can arise. From an economic point of view, either strategy leaving the manifold $S_r = S_s > 0$ is optimal as the comparison yields the same value. The resolution of the question of maintaining the condition is provided by the Generalized Legendre-Clebsch Condition.

Thus the dispersal surface is the bearer of nonuniqueness in the optimal solution to the capital-budgeting problem. This condition has its analog in the static analyses of capital budgeting in the following sense. In most static formulations which include internal as well as external equity financing, the assumption is implicit that either form of financing is equally advantageous in that no special preference for one is obvious. This condition corresponds to, in the context of this analysis, the firm being on the manifold $S_r = S_s > 0$. Thus statically all controls are optimal and this situation yields the prevalent indeterminant solution found in static analyses. Dynamically, the situation has been reduced to nonuniqueness since there are two equally valid optimal policies to follow.

There will be no unique trajectory of capital expansion reflecting equilibrium as the planning horizon approaches infinity. However, there can be two trajectories which closely approximate this condition of constant price-equity ratio and constant dividend payout ratio. This condition yields a price appreciation equal to dividend growth, with both being constant proportions of asset growth. This condition reflects equilibrium in the capital market. Here again the nonuniqueness of the capital-budgeting firm valuation solution is reflected.

The necessary conditions for optimal control can be interpreted as providing an economic equilibrium condition in the following sense. It has been shown that maximizing the Hamiltonian can be interpreted as maximizing the implicit value of current share ownership. If the shareowner is not at an economic equilibrium point by this maximization, then he would have wanted more or less equity and/or price change than the maximization provided. To assure that this implicit value of share ownership cannot be increased by a shift in equity or price, it is necessary that the derivative of the maximal Hamiltonian with respect to each state variable be zero. Doing this operation, the adjoint variables are again interpreted in accordance with

$$\lambda_P{}^* = \frac{\partial V^*}{\partial P}$$

and

$$\lambda_E{}^* = \frac{\partial V^*}{\partial E}.$$

Under the conditions of this problem, i.e., the optimal controls are constant and assuming that the second partial of V exists, the necessary conditions for the Hamiltonian to be maximal in terms of changes in the state variables are

$$\frac{d}{dt}\left[\frac{\partial V^*}{\partial E}\right] + \frac{\partial H}{\partial E} = 0$$

$$\frac{d}{dt}\left[\frac{\partial V^*}{\partial P}\right] + \frac{\partial H}{\partial P} = 0.$$

Note, however, that these equations are the Euler-Lagrange equations of the necessary conditions for optimal control [see section 5]. Thus the necessary conditions of optimal control require the implicit value of share ownership to be maximal and the shareholder to be at an economic equilibrium point at every instant of time.

8. An analysis of rate of return

■ The main purpose of this paper has been to provide a framework to show the effect of a rate-of-return specification on the structure and composition of long-range investment decisions and the valuation of a utility. The results clearly demonstrate how dramatically this specification can affect these decisions. For relatively small changes in rate of return, totally different investment programs become possible and optimal. This impact of small changes in rate of return on capital budgeting decisions and utility valuation has long been observed empirically, and this paper provides one theoretical framework from which to analyze possible consequences quantitatively. Two observations on the solution cases warrant special attention.

Consider cases 1 through 3, where the rate of return to equity is either less than or very close to the investor discount rate. The analysis shows that, in the main, the optimal policy will be to pay out all earnings in dividends and never re-invest internally. External investment is only desirable as long as a transient favorable price-equity ratio exists. This result is consistent with economic theory, since it is well known in financial analysis that if the internal rate of return for an investment is less than the external or market rate of return for investments of comparable risk, then an alternative investment is more desirable. An alternative name for ρ, the investor discount rate, is the market rate of return for investments of comparable risk. These solution cases represent the situation of a "non-growth" return on capital. If sustained growth of the utility is to be encouraged, then the rate of return must be set above this "bare bones" level. One interpretation of this is that a "fair" rate of return must satisfy at least the inequality

$$r > \frac{\rho}{1 - \delta},$$

where the "bare bones" level would be defined as

$$r = \frac{\rho}{1 - \delta}.$$

Solution cases 4 and 5 seem to characteristically represent the central range of rates of return. It has long been recognized that a fair rate of return to equity capital must reflect in some manner the growth requirements of a utility from a viewpoint of capital attraction capability [17]. Note that this factor, k, is explicitly included in cases 4 and 5 and actually determines in part the defining range of rate of return. The nonuniqueness of the capital-budgeting solution appears in these cases. Also, any rate-of-return specification which acknowledges continuous growth considerations will be found here.

It is quite reasonable to interpret k as reflecting the rate of investment-need as supported by some aggregate demand for service. One way of ensuring a consistent growth at rate k over the planning horizon is to set

$$r = \frac{\rho}{1 - \delta} + k\left[1 + \frac{1}{c(1 - \delta)}\right].$$

This is the upper bound for solution case 4 and effectively prevents the utility from entering region B of case 4, the "non-growth" area, for characteristic initial starting positions. It also forces a price-equity ratio close to $1/1 - \delta$, a relationship which has some empirical support.

Notice that this rate of return is a sum of the "bare bones" level plus a factor which reflects the growth requirements. Thus if no growth is warranted, i.e., $k = 0$, the "bare bones" static case results. Also, this rate of return reflects the need for rate-of-return adjustments as economic conditions change, as reflected by ρ for changes in "riskiness" and k for changes in growth requirements. This rate of return seems to satisfy the "fair" rate-of-return requirements [17]:

1. comparable earnings standard for investments of comparable risk, (ρ),
2. attraction of capital standard to reflect growth requirements, (k).

These two standards have arisen principally from a recognition that utilities must compete for capital and must do so more or less constantly. The model and analysis in this paper furnish a way of reflecting these standards explicitly in a rate-of-return relationship.

9. Conclusion

■ The model demonstrated here allows rate of return to be specified by prudent measurement of variables of recognized significance. What, if anything, can be said about the relative merits of this model and others mentioned earlier? The answer lies in comparing the merits of the underlying assumptions. Each model is a simplified representation of a very complex process. The model developed here has several appealing characteristics in terms of parameter measurability and dynamic behavior. But it also contains several strong behavioral assumptions, especially those concerning the capital market. This tradeoff is characteristic of models of rate of return under conditions of growth, since the ingredients required for determining rates can only be quantified either by imposing behavioral assumptions or by using uncertain parameters (in terms of measurability).

The results of this model and analysis suggest that there will be interesting differences between dynamic financial analyses of the firm and the familiar dynamic analyses in economics, e.g. macro growth models. The model and analysis here have indicated some of these differences.

Growth analyses characteristically use a closed macroeconomic system; the model here is a micro financial model.

The model here did not yield the familiar singular equilibrium solution (analogous to a "golden path"). This could have been remedied easily by changing the objective functional; however, the objective functional was chosen as the traditional representation of value of share ownership. A departure from this traditional representation in order to arrive at an equilibrium solution might be construed as suggesting that traditional theory precludes one.

The present model also produced a dispersal surface (in some of the solution cases), which has not been the case in any growth model. This dispersal surface has a meaningful economic interpretation, while a meaningful counterpart in a macro model is difficult to imagine.

These differences seem interesting and suggest a rich area for research, not only for applications in regulatory theory but in the general area of financial growth and cost of capital as well.

Appendix

■ Solution synthesis is a difficult part of solving optimal control problems, particularly nonlinear ones. The approach discussed in section 5 is a systematic method for attacking this step. An abbreviated sketch of this approach, drawn from [5], is given below.

Recall the statement of necessary conditions on page 257. To these are adjoined terminal conditions on the state variables in terms of a parameterization of the terminal manifold in the state space. Let

$$P(T) = s_p > 0$$

$$E(T) = s_E > 0 .$$

Define

$$\tau = T - t .$$

Thus τ represents the time needed to reach T. Since T is assumed to be a known constant, let

$$\mathring{y} = \frac{dy}{d\tau} .$$

It follows that

$$\mathring{y} = - \dot{y} ,$$

and the state and adjoint equations (equations 1, 2, 3, and 4 on page 257) in the reverse time change in sign. Initial conditions (in this reverse time sense) for these equations are provided by the tranversality condition and the parameterization of the terminal manifold, i.e.,

$$\lambda_P(0) = 1 \quad P(0) = s_P > 0$$

$$\lambda_E(0) = 0 \quad E(0) = s_E > 0 .$$

To start the construction, the Hamiltonian must be maximized at $T(\tau = 0)$. To do this, S_r and S_s must be known (recall the S_i are the switching functions defined on page 257).

At $\tau = 0$

$$S_r = Er[\lambda_E - c\lambda_P - 1] = - s_E r(c + 1) < 0$$

$$S_s = Er\lambda_E \left[1 - \frac{E}{(1 - \delta)P} \right] = 0 .$$

Thus initially $S_r < 0$, $S_s = 0$, and this singular condition must be examined.

$$S_r < 0 , \; S_s = 0 \;\; \text{implies} \;\; u_r^* = 0 , \; 0 \leq u_s^* \leq \frac{k}{r} .$$

$$S_s = 0 \text{ implies } Er\lambda_E \left[1 - \frac{E}{(1 - \delta)P} \right] = 0 .$$

Since $S_r < 0$, then $Er \neq 0$ and this requires

$$\lambda_E \left[1 - \frac{E}{(1 - \delta)P} \right] = 0 .$$

For this equation to be satisfied for a non-zero time interval implies that all time derivatives must vanish. There are three subcases:

1. $\lambda_E = 0 , \quad \left[1 - \frac{E}{(1 - \delta)P} \right] \neq 0$
2. $\lambda_E \neq 0 , \quad \left[1 - \frac{E}{(1 - \delta)P} \right] = 0$
3. $\lambda_E = 0 , \quad \left[1 - \frac{E}{(1 - \delta)P} \right] = 0 .$

Suppose $\lambda_E = 0$, then by equation (2) it is required that

$$\lambda_P = -\frac{1}{c} \text{ a nonzero constant} .$$

However, if $\lambda_E = 0$ there is no solution of equation (1) which satisfies the above relationship. Thus the condition $\lambda_E = 0$ cannot be satisfied for a non-zero time interval. For S_s to be zero, it is necessary that

$$1 - \frac{E}{(1 - \delta)P} = 0$$

for a non-zero time interval. Differentiation yields

$$Er = \rho P .$$

Further differentiation yields the same results. Thus it is required that

$$\frac{r}{\rho} = \frac{1}{1 - \delta} = \frac{s_P}{s_E}$$

for this condition to be sustained. Consider the case where

$$\frac{r}{\rho} > \frac{1}{1 - \delta} ,$$

hence $S_r < 0$, $S_s = 0$, holds only on the terminal manifold. It is necessary to examine the derivative of S_s at $\tau = 0$ to determine the sign of S_s at $\tau = 0^+$. The result [5] will be that

$$S_s(0^+) < 0 \quad \text{if} \quad \frac{s_P}{s_E} \leq \frac{1}{1 - \delta} ,$$

$$S_s(0^+) > 0 \quad \text{if} \quad \frac{s_P}{s_E} > \frac{1}{1 - \delta} .$$

Consider the case where $s_P/s_E \leq 1/1 - \delta$. Then $S_r(0^+) < 0$, $S_s(0^+) < 0$ and $u_r^* = u_s^* = 0$ are the extremal controls. The state equations become

$$\mathring{P} = c[\rho P - Er]$$
$$\mathring{E} = 0\,,$$

and the adjoint variables are defined by

$$\mathring{\lambda}_P = -(c+1)\rho\lambda_P\,,$$
$$\mathring{\lambda}_E = rc\lambda_P + r - \rho\lambda_E\,.$$

Their solution yields

$$\frac{P}{E} = \left[\frac{s_P}{s_E} - \frac{r}{\rho}\right]\exp[\rho c\tau] + \frac{r}{\rho}$$

and

$$\lambda_P = \exp[-(c+1)\rho\tau]$$

$$\lambda_E = \frac{r}{\rho} - \frac{r}{\rho}\exp[-(c+1)\rho\tau]\,.$$

Since $r/\rho > 1/1 - \delta$, P/E decreases with τ. Using these solutions, S_r and S_s can be evaluated. It follows directly that S_s will remain negative as long as $S_r < 0$ and $P/E < 1/1 - \delta$. Evaluating S_r yields

$$S_r = Er\left\{\frac{r}{\rho} - 1 - \left\{\frac{r}{\rho} + c\right\}\exp[-(c+1)\rho\tau]\right\}\,.$$

This will be zero when

$$\frac{r}{\rho} - 1 = \left\{\frac{r}{\rho} + c\right\}\exp[-(c+1)\rho\tau]$$

or when

$$\hat{\tau} = \frac{1}{\rho(c+1)}\ln\frac{\rho c + r}{r - \rho}\,.$$

Now at $\hat{\tau}$, $S_r = 0$, $S_s < 0$, and since this is a singular situation it must be checked in the manner followed on the terminal manifold. By proceeding with this construction, the full solution can be gained. To do this here would be lengthy and outside the scope of this paper. Considerable additional analysis is required, particularly on the manifold $S_r = S_s > 0$, and is available in full detail in [5].

References

1. Arrow, K. J. "Applications of Control Theory To Economic Growth," *Technical Report No. 2*, Institute of Mathematical Studies in the Social Sciences, Stanford University, 1967.
2. Averch, H. and Johnson, L. L. "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, Vol. 52 (December 1962), pp. 1053–1069.
3. Bellman, R. *Dynamic Programming*. Princeton: Princeton University Press, 1957.
4. Colbert, A. R. "Common Stock Earnings Allowance for Utilities as Adjusted for Growth," *Land Economics* (February 1963).

5. Davis, B. E. and Elzinga, D. J. "Solutions of an Optimal Control Problem in Financial Modeling." Forthcoming.
6. Dobell, A. R. and Ho, Y. C. "Optimal Investment Policy: an Example of a Control Problem in Economic Theory." *IEEE Transactions on Automatic Control*, Vol. AC-12 (February 1967).
7. Gordon, M. J. *Testimony of M. J. Gordon*, F.C.C. Staff Exhibit No. 17, Federal Communications Commission Docket 16258.
8. ———*The Investment, Financing, and Valuation of the Corporation.* Homewood, Illinois: Richard D. Irwin, 1962.
Homewood, Illinois: Richard D. Irwin, 1962.
9. ———, and Shapiro, E. "Capital Equipment Analysis: The Required Rate of Profit," *Management Science* (October 1956).
10. Hestenes, M. R. *Calculus of Variations and Optimal Control Theory.* New York: Wiley, 1966.
11. Isaacs, R. *Differential Games.* New York: Wiley, 1965.
12. Johnson, C. D. "Singular Solutions in Problems of Optimal Control," in *Advances in Control Systems: Theory and Applications*, vol. 2. New York: Academic Press, 1964.
13. ———, and Gibson, J. E. "Singular Solutions in Problems of Optimal Control," *IEEE Transactions on Automatic Control*, Vol. AC-8 (January 1963).
14. Kelley, H. J. "Second Variation Test for Singular Extremals," AIAAJ, Vol. 2 (1964).
15. Lee, E. B. and Markus, L. *Foundations of Optimal Control Theory.* New York: Wiley, 1967.
16. Lerner, E. M. and Carleton, W. T. "The Integration of Capital Budgeting and Stock Valuation," *American Economic Review*, Vol. 54 (September 1964).
17. Leventhal, H. "Vitality of the Comparable Earnings Standard for Regulation of Utilities in a Growth Economy," *Yale Law Journal*, Vol. 74, No. 6 (May 1965).
18. Lindsay, R. and Sametz, A. *Financial Management: An Analytical Approach.* Homewood, Illinois: Richard D. Irwin, 1963.
19. Miller, M. H. "The Corporation Income Tax and Corporate Financial Policy," in *Stabilization Policies*, Englewood Cliffs, N. J.: Prentice-Hall, 1964.
20. ——— and Modigliani, F. "Dividend Policy, Growth, and Valuation of Shares," *The Journal of Business* (October 1961).
21. Pontryagin, L. S., et al. *The Mathematical Theory of Optimal Control Processes.* New York: Wiley, 1962.
22. Sparrow, F. T. "Mathematical Models of Growth Allowance for Public Utility Regulation: A Synthesis," *Management Science*, Vol. 13 (February 1967).
23. Tasch, P. Ph.D. dissertation, Johns Hopkins University, 1969.
24. Westfield, F. M. "Regulation and Conspiracy," *American Economic Review*, Vol. 55 (June 1965), pp. 424–43.
25. Williams, J. B. *The Theory of Investment Value.* North Holland Publishing Company, 1938.

The effectiveness of the Federal Power Commission

Paul W. MacAvoy

Professor of Management
Massachusetts Institute of Technology

This paper takes the view that the Federal Power Commission dispenses services that have measurable economic benefits and imposes the costs of these services on both the regulated firms and the final consumers of gas and electricity. An attempt is made to define and measure benefits from regulation at the margin, where this margin has been chosen by the Commission via present rulemaking and surveillance activities. The costs of regulatory proceedings are estimated to include expenditures of the Federal Power Commission and other participants in the Commission's proceedings, and to include implied losses of final consumers consequent from regulatory delay. Benefits are compared to costs for each of the Commission's areas of responsibility, and the comparisons pose the question whether there ought to be more or less regulatory activity. The estimates here imply that the F.P.C. is operating at a greater scale than net benefits warrant, particularly as a consequence of its ventures toward regulating natural gas production in the last decade.

■ The Federal Power Commission can be credited with outstanding administrative performance in the past decade. Long known as the independent regulatory commission with the highest standards for legal procedures and rulemaking, the Commission had become paralyzed by its own requirements for detailed analysis and review when faced with setting prices for thousands of natural gas field sales each year. Early in the 1960's, however, the Commission invented the means for clearing out the natural gas cases, and by May of 1968 the Supreme Court had affirmed this "area rate" process of reviewing large numbers of gas field sales in a single case or docket. At the same time, wider jurisdiction had been claimed by the Commission over wholesale distribution of electricity; in the words of one electricity company executive, "the Commission, under Chairmen Swidler and White, greatly extended the concept of what constituted an interstate sale . . . [while] many state Commissioners still assert[ed] duplicate jurisdiction over such sales."[1] Here again

---

P. W. MacAvoy holds the A.B. degree from Bates College (1955), the M.A. from Yale (1956), and the Ph.D. from Yale (1960). He was Assistant Professor of Economics in the University of Chicago's Graduate School of Business from 1960 to 1963. After coming to M.I.T. in 1963, he joined the Sloan School of Management in 1966 and was appointed Professor of Management in 1969. Professor MacAvoy is Editor of THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE.

This is the first of a series of papers leading to a comprehensive systems analysis of the Federal Power Commission. The project is being supported by the Brookings Institution under its program of Studies in the Regulation of Economic Activity, which is financed by a grant from the Ford Foundation. The author expresses his gratitude for their support, without holding them responsible for the results. He is also grateful for the suggestions and criticisms of M. A. Adelman, Paul Bradley, Robert Crandall, Edward Erickson, Clark Hawkins, Phyllis Klein, Elliott Morss, Stewart Myers, Merton J. Peck, Richard Posner, Richard Schmalensee, Phillip Sporn, and Haskell Wald.

[1] Private correspondence dated November 18, 1969.

the Commission made findings in cases that set standards for due process under new and extended regulation.

While the Commission has managed to get new types of dockets completed, it has not been able entirely to dispel concern over the effectiveness of regulation. The new decade has begun with prognostications of shortage of natural gas, not only in the trade magazines but also in a report from the Commission's own Bureau of Natural Gas. [21] The electric power producers now lack the capacity to provide peaking service in a number of major population centers that had been accustomed to such service. Regulation has been given part of the responsibility for such conditions in these industries.[2] At least the possibility exists that the Power Commission has devised admirable procedures for achieving economically deplorable results.

This essay begins the investigation of such a possibility. The Commission is viewed as an organization using resources and commanding their use by others for the purpose of changing incomes and outputs in electricity and gas markets. The time frame of reference is a single fiscal year in the late 1960's, so that—except in the cases of "new" gas field and electricity regulation—the costs and benefits are from one short-run period of ongoing activities of the agency. The analytical frame of reference is the concept of efficiency—of gaining the most in services or outputs desired by consumers for expenditures on regulation.

These are rather narrow frameworks. To determine whether the Commission could have effectively spent more or less in the last fiscal year is not to document the gains from regulation in the long run. The reason for doing this limited study is that quantitative evidence can be brought to bear on the question of whether there would be net gains from a larger or smaller Federal Power Commission. The evidence that would make it possible to judge the economic gains from having any Commission at all remains to be discovered.

The last section of this essay contains some conclusions on whether the "scale" of regulatory activities in gas or wholesale electricity is "overextended." Tentative proposals are made for changing procedures or criteria for regulating prices so as to reduce the costs of regulation or to add to consumer gains at present costs. But these are tentative because Commission rules are changing. Decisions are not based on standards that are consistent or discernible. Even if the Commission is erratic in its decisions, however, these conclusions are important since they are based on the latest Commission decisions. In that case, any reversal in later decisions would have implications in direct contrast to those of past decisions.

Where this exercise does lead to proposals for economic reform, the question remains as to what exactly should be done. If the Commission is "inefficient" in an economic sense, it does not necessarily follow that Congress or the Commission should make changes. Full requirements for economic efficiency could not possibly be put into effect, since the Commission was not brought into the world cast in such a mold nor was it given a mandate to acquire what it had not inherited. But there might be little objection to making changes in Commission procedures which *add to* economic efficiency, as long as these do not restrict the legal rights and prerogatives of the parties to regulation. This argument at least is one source of optimism in the face of foreboding forecasts for conditions in the electricity and gas industries in the 1970's.

---

[2] At least accusations are made in the magazines, if not the publications of the F.P.C.'s Bureau of Natural Gas. Cf. Liversidge [7] and Main [11]. Cf. also Kitch [6]; at p. 278 the author concludes: "In the field market the regulation will only cause the loss of gas which would otherwise be produced. In the consumer markets the possible consequences are even more disturbing."

1. Efficiency in regulation

■ The independent regulatory commission has a mandate from Congress and the Federal courts to provide a complex set of services to the industry with which it is concerned, and ultimately to consumers of the final products of that industry. The commission could be found to be "efficient," in the technical sense of the term, if there were no alternative way of providing the given regulatory services at lower annual expenditures of Federal funds. But this is a narrow definition of efficiency, since the possibility exists that the Commission is minimizing purely legal and administrative costs of promulgating an unwanted or unnecessary set of rules.

The direct costs of the Commission are not the total costs of regulation. The Commission at any time could show reduced expenses by rulemaking that required the companies being regulated to bear additional costs, or that required final consumers to bring cases. The Commission could spend a great deal less by carrying out its usual activities at a slower rate. Both possibilities show that the expenses of all parties involved in regulatory proceedings should be considered the observable "costs of regulation" in any well-defined time frame of reference. There are indirect costs—especially involved in delay—which are reflected in higher prices for final goods and services. They should be included in the final balance sheet for the Commission.

The activities of the Commission are supposed to prove beneficial to the consumer and to the economy as a whole. Commissions in general are organizations for limiting business decisions, or for preventing the occurrence of certain patterns of market behavior in particular industries. The Commission's service is the nonoccurrence of particular levels of price or particular quantities of products. It could be involved as well in designating the sources and amounts of products. The benefits to society can be measured by the difference between the value under regulation of the economy's goods and services and that value when the prohibited patterns occur. There are all kinds of measures of value, but the measure allowing comparison with expenditures is the dollar "surplus"—that amount of money which consumers and producers together would be willing to pay over and above the actual amount involved in transactions in order to maintain the conditions from regulation. Finding this in a particular case is difficult. But measuring the surplus by areas under demand functions $q = f(P)$ for quantity q and price P on final products has elsewhere succeeded in providing an estimate of value for policy analysis. The Commission is efficient if it is extending services at least cost and if the services result in dollar consumer and producer surplus greater than these costs.

The Federal Power Commission's day-to-day activities are very loosely organized around a few long-established public policies, given in the *Federal Power Act* of 1935 and the *Natural Gas Act* of

1938 and restated by the courts in thirty years of precedent decisions. One policy is protection of consumers against prices that are "unreasonable" and "unduly discriminatory."[3] To put this policy into effect on wholesale distribution of power and natural gas, the Commission has generally required that the company's average prices be equal to average "costs of service," however these costs on wholesale service are discovered in company cost accounting records. Even more exactly, prices for sales in a specific contract ought to be set close to costs of service, because the "cost of service formula is the one best adapted in determining just and reasonable rates in an industry where costs are known."[4]

Protection of consumers from price-cost disparities has not been the only policy in regulation. The Federal Power Commission was given the task of "assuring an abundant supply of electric energy throughout the United States with the greatest possible economy and with regard to the proper utilization and conservation of natural resources" in Section 202 of the *Federal Power Act.* Public intervention to provide more service at higher quality was even more direct in the *Natural Gas Act*; here the Commission can and should require improvement and extension of a pipeline's service when it is in the public interest. The emphasis on quality has in fact appeared in F.P.C. certification of new facilities for producing electricity and transporting gas—here detailed reviews are made of the entrant's financial resources and fuel capacity to "meet the demands which it is reasonable to expect will arise."[5] Also, the Commission has become involved in the planning and forecasting processes in these industries, to a very limited extent "encouraging" coordination in the operation or growth of larger systems.

The "price" and "quality" policies have been put into effect in the Commission's ongoing surveillance and case activities. The price reductions captured by a year's work can be characterized by a movement from P_1 to P_2 in Figure 1; that is, buyers in the absence of the Commission's ongoing surveillance activities would find prices "creeping up" to P_1 or prices "failing to be reduced" to P_2, but with a year's regulatory vigilance (and expenditure on surveillance and litigation) price would necessarily be at P_2.[6] The gains for the consumers from regulation are shown as $(P_1 - P_2)Q_1$, where Q_1 is the amount of service demanded at the higher price; at least this is all that has been claimed by the Power Commission on interstate gas and wholesale electricity sales. But these gains in area "*A*" are not the only possible benefits from regulation.

---

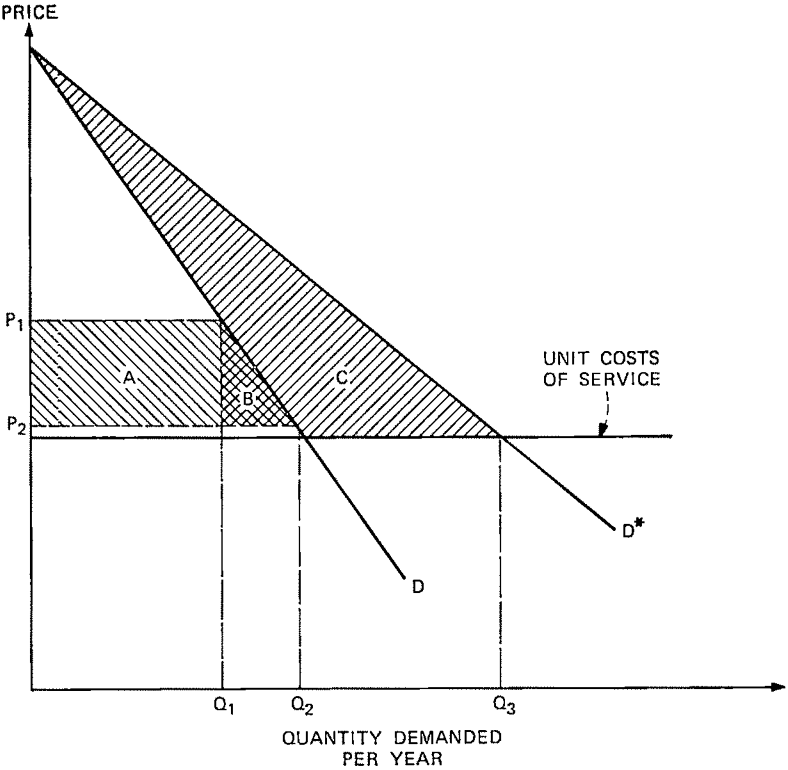
[3] As shown by the more recent court decisions Cf *F P C. vs Southern California Edison Company* 376 U. S. 205 (1964) on the jurisdiction and standards of the Commission for electricity prices, and *F.P.C vs Northern Natural Gas Company* 346 U. S. 922 (1954) for "costs of service" and prices of natural gas pipeline companies.

[4] *In the matter of South Carolina Generating Company* 16 F.P.C. 52 (1956) at 58. The qualifying clause "where costs are known" is important, in the many instances when there are widely varying estimates of costs, prices are not set equal to one of these estimates. The best characterization of the general relation is that "costs of service" provide a lower bound for prices suspected of being "too low" and discriminatory, and a much more approximate upper bound for prices found to be "too high."

[5] Cf. *In the matter of Kansas Pipeline and Gas Company and North Dakota Consumers Gas Company* 2 F.P.C. 29 (1939).

[6] The price would presumably be higher than this representative P_1 if the Commission were abolished, however It is assumed here that the Commission is in operation, but at a much reduced level of expenditures on compliance, when P_1 is in effect In fact, the upper limit of P_1 would be a price level clearly in violation of the precedents set by previous F.P C court cases dealing with that industry, so that a case without surveillance information could be brought if that level were exceeded

FIGURE 1



Given that consumers' demands are represented by the negatively sloped demand curve *D*, the Commission-induced price reductions increase the quantity demanded from Q_1 to Q_2. This uncovers a second consumers' gain, shown as area "*B*," that is not derived from income redistribution. This is the dollar equivalent to what consumers would pay for $(Q_2 - Q_1)$ over and above actual payment $P_2(Q_2 - Q_1)$ rather than go back to price P_1 and quantity Q_1.[7] As such, it is an increase in the "quality of service" derived from direct price regulation and should accompany the area "*A*" benefits from "protection of the consumer."

The direct pursuit of "quality" is not by means of price reductions, however. Increasing the accessibility and reliability of service, by regulating the form and scale of capacity additions, should shift out demand from *D* to *D**. This uncovers consumers' surplus, shown by the area "*C*," and increases the quantity demanded to Q_3. Such gains may be had only at increased "costs of service," however, so that they are available only at prices higher than P_2. They not only do not accompany reductions in price, but in some cases are not realized unless price is maintained at the higher level.[8]

The critical question is whether any of these measures can actually be used to make the case for the regulatory *status quo* or alternatively for regulatory reform. In the representative regulatory commission, the information produced in the regulatory process is not for the purpose of making such assessments and is not useful for doing so. This is true for both costs and gains.

---

[7] See Winch [23] These gains are not to be realized unless certain conditions prevail. First, prices have to be greater than the total economic costs of the additional output. The measured "costs of service" are not marginal costs, except by chance equality of accounting overhead allocation with the opportunity costs of using fixed factors in this production. There is no way of making certain that this condition is met by recourse to regulatory accounting information. A second qualifying condition is as important the $\triangle P \triangle Q/2$ measure is a good measure of aggregate gains of consumers only if their distribution to some and not to others is judged to be acceptable Here every dollar is taken to be of equal social value, no matter who receives it—the steel company using gas or electricity for production, or the poor widow-orphan using it for light and heat.

[8] The analytical basis for this contradiction is shown in Schmalensee [13].

The costs of commission operations are calculated in straightforward accounting fashion in the budget and expenditure review departments of the Federal Government. But no assessment is made of costs imposed on the regulated firms—the costs of defense or intervention in the regulatory proceedings.

Estimates of benefits depend critically on detailed information on market prices and quantities in the absence as well as the presence of the ongoing regulatory surveillance and accounting activities. Often there is some information on market behavior before and after some surveillance activity took place. A good part of this information is ambiguous because it is not sufficiently detailed to allow separation of increases in demand as the market grows from changes in quantity demanded (where the latter follow from price reductions resulting from this activity). Also, prices after regulatory surveillance are related to "costs of service." The calculation of "costs of service" of the regulated companies is so arbitrary that it is not clear whether finding them leads to price reductions. The basic problem in the calculation is the allocation of joint capital expenditures among recipients of different kinds of service; there are no rules by which the resulting unit costs can be made to approximate the economic costs on which unregulated prices will be set. In general the calculations by the Commission or by producing firms of the "costs of service"—equal to variable costs plus an allocation of undepreciated capital expenditures—may approximate unregulated prices just as easily as they may approximate long-term economic costs so as to lead to lower regulated prices. Last of all, the information on "quality" of service is in most cases limited to quantity information, or to recordings of percentage "down-time" of systems. Nowhere are accessibility and reliability assessed in terms of final consumer demand.

Even though near impossible, an initial attempt to assess benefits and costs will still be made in the case of the Federal Power Commission's regular activities. The costs of regulation can be estimated from Commission expenses and from a first limited survey of the defendant's actual expenditures in recent cases. The benefits are assessed from the Commission's own public portrayal of its accomplishments. Wherever the F.P.C. shows a dollar of benefits that can be classified under "consumer protection" or "improved quality" of service, it may be questioned and then reduced but not denied. This procedure results in an optimistic estimate of Commission benefits minus costs on its ongoing operations.

2. The costs of regulation by the federal power commission

■ The Federal Power Commission's powers to regulate the companies producing or transporting natural gas and electricity are broad, given the general wording in its statutes and the Federal Courts' willingness to let all matters of price and quality pass before the Commission so long as "due process" is observed. But the regulatory process for holding prices to "costs of service" makes up the major part of Federal Power Commission activities each year. The chronological steps in this process frequently begin with an application for a "Certificate of Necessity and Convenience" by the company seeking to introduce the regulated product or service into some new market. The Certificate application is examined in hearings for "sufficiency" of demand for the proposed service and "capacity" of the company to provide the service. This is to assure the quality of service. After certification has been granted and service has begun, the reviews of prices are undertaken first by a hearing examiner, then the Commission, and finally on appeal by the United States Courts. The limit on average price is set close to the Commission's estimate of the costs of capital and variable inputs needed to maintain the service demanded. Finally, there are surveillance procedures for ascertaining that the price-profit ceilings go into effect. All of these reviews require expenditures on Commission, court, and company personnel and supporting resources.

The expenditures of the Commission varied from $11 million to more than $14 million during fiscal years 1964 to 1968. Part of this outlay was for services having market value and not related to regulation, however. In particular, some were incurred in managing resources on Federal lands, Indian lands, or in providing maritime navigation within the jurisdiction of Federal power facilities. These outlays, along with those on surveillance of the activities of other Federal agencies in the hydroelectric power business, cost $3.8 million in the fiscal year 1968.[9] The regulatory outlays incurred by the Commission, or debited to firms by the Commission, were the difference between gross expenditures and these resource management expenditures, or approximately $10.6 million in the last recorded fiscal year.

This amount was a small part of the total costs to the economy of a year's regulatory proceedings, however. Most of the costs of regulation were incurred by the producers of gas and electricity or by the transportation companies in these industries that were defendants in cases or the subject of "cost of service" reviews. There were roughly six classes of proceedings in a year which required the services of industry experts, lawyers, and corporate executives for extended periods of time. The cost of these resources varied from case to case, but estimates can be constructed of average costs for each type of proceeding which are representative of most of the instances. In some cases, actual costs for all of that type can be calculated. The total defendants' expenditures perhaps come to more than $36 million in a representative fiscal year. This amount might seem much larger than expected, but a category-by-category review suggests that it is not unreasonable.

☐ **The costs of the electric power regulatory activities of the commission.** After thirty years of leaving regulation to the States, the Commission has carried out a campaign to set wholesale ceiling prices on all sales in interstate commerce since a 1964 Court decision affirming regulatory jurisdiction.[10] The catching-up process has centered activities on reviewing existing price schedules, rather than first certifying service and then reviewing prices. Approximately 2200 schedules have been filed each year with supporting cost information. The Commission, on receipt of the filings, has reviewed virtually all of them. A minority are the subject of the 50 "rate" or price studies carried out each year, and also the basis for the 30 actual cases on "rates, terms, and conditions" brought before the entire Commission and appealed to the courts.

---

[9] See [22], Appendix, p. 930.

[10] Cf. [17] for 1964, pp. 85 and 165.

Each step of this procedure is costly. The Federal Power Commission itself spent $1.6 million in fical year 1968 on electric power utilities regulation (according to the Budget of the United States Government). Discussions and correspondence with a number of power companies indicates that, even if the filing of schedules involves no new rate or costing decisions, they cost anywhere from $25 to $1000 of legal services and computer time to prepare. The total costs of more than 2000 filings by the companies each year can be conservatively estimated at $50,000.

The rate studies and cases initiated by the Federal Power Commission are much more expensive. The Commission staff initiates the studies informally as part of annual surveillance procedures when the first review of rate schedules and annual reports of costs suggests that a more complete analysis is warranted. Their requests for detailed capital expenditures and output patterns are so extensive that some part of normal employee time in the company has to be devoted to this regulatory work. With 53 rate studies in progress in the 1968 fiscal year, most involving new calculations of "costs of service," the total indirect costs of this stage of regulation to the companies must have been extensive. A number of power companies show "charges related to filings" of more than $1000 which are incurred while filling requests for information.[11] This stage in the annual regulatory review must cost the industry $50,000 per annum.

If the informal price review does not lead to the reductions sought by the Commission, then a formal rate case is initiated. Approximately 30 cases are in process each year, and 10 of these are usually completed. In the three years a case is in process, a full "cost of service" study is completed by both company and Commission, pre-hearing conferences are held, hearings are undertaken, and finally a Commission decision is rendered. All these steps involve investigations and testimony by company cost experts and executives, as well as the services of accountants, economists, and lawyers. A review of the 29 cases shown in the 1968 Federal Power Commission *Annual Report* shows that the full company costs for proceedings "in docket" that year ranged from $7000 for a review of a single rate in Maryland to $305,000 for a full scale rate review on service to a group of cooperatives in North Carolina. Most cases involved additional costs of delay for the defendants and of uncertainty as to future charges and costs of service for both parties. Then the 10 cases finished in a year must cost at least $1,500,000 when all costs over the case lifetime are included. The total costs of regulating electricity wholesale prices probably comes to an estimated $3.2 million per year when both Commission and defendants' expenses are included.[12]

The nature of the cases themselves points to the possible occurrence of further indirect costs of regulation. A large number involve disputes between small cooperative companies retailing electricity and large power-producing networks; the point of dispute is the necessity for the wholesale price to be equal to that charged some other buyer of the same or different size at some other location.[13] These disputes do little more than demonstrate to the power producers that, in order to avoid hundreds of thousands of dollars of litigation, they must offer the small buyers prices reflecting economies of scale which do not exist for those buyers. This condition must provide incentives not to connect to the small buyers—at the indirect costs of higher generation expenses for locally produced electricity.

---

[11] Cf. [18] for 1968 under the accounts "regulatory expense" (353) and "consultants expense" (355).

[12] This does not account for the possibility of lower or higher costs in any one year. If all cases were "single-rate" cases, costs for defense might be as low as $100,000, but if all were complex interclass rate reviews, then defense costs might be $3,000,000.

The Federal Power Commission has embarked most tentatively to add to the "quality" of service by initiating cases requiring the larger producers to provide power to small retailers. To this point, the campaign has had little effect except on the defense costs of regulation. The most promising docket was E-7257, in which the city of Gainesville, Florida, requested the F.P.C. to require interconnection with Florida Power Corporation, and the F.P.C. examiner approved the request along with a procedure for sharing the cost savings in electricity generation. The Commission reversed the examiner, and, in the fourth year of this proceeding, the city of Gainesville's plea is before the circuit court at New Orleans. The costs of this and the few other "extension of service" cases have been a little more than $100,000 apiece so far; if only two are settled in a Commission working year, then the costs per annum of this phase of regulation may be $200,000.

After prices have been set within the framework of regulation, the day-to-day operations of electric companies take place according to rules set once again by the Commission. Capital equipment used in operations has to be that specified in "costs" that determine rates. The Power Commission has been conducting field examinations of those capital expenditures actually made by companies to ascertain that they are the "original costs" for providing the regulated services. More than 200 electric utilities filed documentation on capital assets actually in use; more than 350 companies filed studies of classification of original capital expenditures among types of service as well. These studies had value to the firms in their day-to-day operations, so that not all of their costs can be attributed to regulation—for one, the companies learn where their equipment is. But the resources devoted to these activities are clearly beyond those found in cost surveillance in unregulated companies, and they have not been justified by the companies as "cost effective" in terms of lost and found equipment. The F.P.C. costs of "electric power industry systems evaluation" were $1.3 million in fiscal year 1968. A small sample of companies shows that their costs on these activities cannot have been less than $5000 per annum in each company undertaking these procedures in any one year.[14] There have been approximately 40 documentation and 75 classification reviews of present or proposed future facilities actually completed in a fiscal year. A minimum estimate of the resource costs of the F.P.C.'s electricity surveillance program must be at least $1.9 million per annum, as a result.

An overall review of the regulation-related costs of the power industry and Commission shows that approximately half are attributable to price control activities and the other half to attempts to improve the quality of service. The companies and Commission together seem to have spent $3.2 million on rate or price regulation in a typical year in the late 1960's, although variations in the costs of the individual cases may have made this amount as high as $4.7 million or as low as $1.8 million. They probably spend $3.4 million per annum on certificate and systems evaluations to extend the quality and quantity of service. The total costs of these aspects of regulation have been in the range from $5 million to $8 million, and have been close to $6.6 million.

---

[13] Cf. E-7129, E-7183, E-7273, E-7308, E-7344, E-7421, and E-7426 in Table 10 of the F.P.C.'s 1968 *Annual Report* [17].

[14] This is a sample average for thirty firms showing "regulatory expense" on Form 1 of the electric power company reports to the Federal Power Commission. These firms did not have under way other proceedings mentioned above, or did not report costs related to system reviews.

☐ **The costs of gas field price regulation.** Regulation of prices charged by producers for natural gas sold to interstate pipelines has been in effect since the Supreme Court decision of 1954 in *Phillips Petroleum Company v. Wisconsin* (347 U. S. 672). But there was not effective regulation, with Commission-imposed limits on prices, during most of the 1960's.

The problem that delayed installation of price control was how to set ceiling prices based on individual company records of "costs of service" when gas comes from wells jointly owned by a number of producers, each with different historical costs, or from different wells in the same market but with histories of different exploration and production expenditures. Faced with having to make arbitrary cost allocations, the Commission moved away from prices based on company "costs of service" to area price ceilings set in reference to country-wide average costs of exploration and development. Provisional prices were set for "new" and "old" gas field contracts in 24 areas in 1960. Formal Commission and Court proceedings have been under way in review of the provisional prices. The Supreme Court in the *Permian Basin Area Rate Case* (390 U. S. 747) affirmed the legality of this price-setting method last year, so that price ceilings can now be set as a matter of course on historical average development costs over large regions.

Regulation of prices is only the most important of a number of Commission activities. The producing company with discovered and developed gas reserves has to go through a number of formal procedures, beginning with obtaining a "Certificate of Necessity" to put the gas into the pipeline soon after a sales contract has been signed. The applications for Certificates now make up an inventory of approximately 2000 per annum, with 1200 reviewed by the Commission in the same year and most of the rest in the following year.

There have been two opposing trends in administering Certificates in the last decade. First, the reviews have become more detailed over time, dealing with contract prices as well as security of supply and demand in the assessment of "necessity and convenience."[15] As a result, filings to the Commission now include a showing that initial prices on new deliveries are "in line" with those in the field region, based on studies of contract pricing and other conditions by lawyers and company accountants. But in contrast to this extension, the Commission has devised a "temporary certificate" allowing extremely rapid installation of new service while the full certificate review is going on. The full Commission certificate reviews take time and hold up the production process, since reserves have to be secured and developed before certification, while production can only take place after the review is complete. The new "temporaries" reduce these costs of delay.

---

[15] That is, since the certificate price can only be reviewed when increased under the present interpretation of the *Natural Gas Act*, this is the only point at which an initial review takes place.

The balance of more complex permanent certificate reviews with substitution of temporary certificates determines the cost of regulatory delay at the present time. In the early 1960's, the costs of delay came to as much as .17¢ per thousand cubic feet for every month of delay and were the reason for a 3.5-percent price premium attached to regulated interstate sales over and above unregulated within-state sales.[16] The delay at that time was more than five months, and if anything the delay at the present time is six months, given that permanent certificates are processed more or less uniformly from one to twelve months after application.[17] But the present delay is ameliorated by temporary certification, so that the price premium required by regulation cannot be half that in the early 1960's. At the most, the temporary certificates could reduce the regulatory delay to one month and the costs of delay to $10 million per annum; at the least, temporaries are worthless because they do not dispel the uncertainty of obtaining the final certificate and compound costs by making delivery possible under a contract that will be rendered null and void. In these worst circumstances, costs have to be those for waiting for a permanent certificate and exceed $58 million for delay.[18]

The step after certification in the Commission's regulation of natural gas is to set producers' prices. As was mentioned, this has been accomplished with maximum "area rates" on all gas in a large geographical region. These rates are then compared with 6000 rate schedules each year to ascertain conformance to the rulings.

The Commission has used or required producers to use far more resources in this pursuit of the ceiling price than might reasonably be expected. The area rate proceedings have been massive and prolonged. The first case began in December 1960 with hearings on prices set by 351 companies under 705 dockets for gas in the Permian Basin of West Texas. Testimony and submissions were completed early in 1964 and were followed by a Commission decision in August 1965 setting a maximum price of 16.5¢ per thousand cubic feet (m.c.f.) on "new" gas-well gas and 14.5¢ per m.c.f. on "old" gas. The Supreme Court affirmed this decision in May 1968, and the Commission then began the last round of putting these ceilings in effect in place of the "interim" or "in line" prices which held at 16¢ on new

---

[16] As documented by Gerwig [5], pp 86–88.

[17] The new gas certified each year has cost

$$C_d = \sum^{n} c_i q_i \left[1 - \frac{1}{(1+r)^d} \right]$$

more as a result of regulatory delay, where c_i is the present value of unit costs of gas quantities q_i in "n" individual contracts, and "r" is the monthly percentage costs of delay during the delay period of length "d." Given other regulatory procedures, c_i has to be not greater than initial base prices set by the Commission for field areas, these prices average close to 17 cents per thousand cubic feet, if it is assumed that contracts were for reserves from new fields discovered in 1967 and 1968. More than 60% of the new fields were in Texas, 10% in Louisiana, and the remaining in other states, weighting the F.P.C. ceiling prices by these percentages results in the $0.17 average. q_i is close to 16 trillion cubic feet, as shown on page 48 of the F.P.C.'s *Annual Report* for the fiscal year 1968 (an anount that has a present value of 6 trillion cubic feet). With delay costs "r" of one percent per month for a "d" of six months, the certification procedure could cost as much as $58 million per year.

[18] Cf. footnote 17 for calculations for the worst case. The most pessimistic estimate of all comes from assuming that the temporaries have zero value—where the producer is indifferent between a temporary and waiting for a permanent certificate. These costs are the costs of waiting.

gas and 11¢ on old gas. The work on a single set of area rates has now almost been completed.

Four more area rate cases have been initiated and are moving along to the decision stage—albeit on a faster schedule, given that the Permian Basin proceedings had provided the rulemaking and even some of the documentation. The approach in each has been the same, and has centered on finding long-term historical average costs of exploration, development, and production. This estimate is used to set future ceiling prices, after adjustments of costs for any special conditions in the area. The search for these costs is burdened with exceptional problems. Foremost is the lack of economic analytical technique to find what is sought: the separable costs of gas alone in the oil-gas regions at both the exploration and developmental stages. Without techniques to center discussion on the quality of data, the actual testimony based on the opinions of those experienced in the industry becomes profuse. The second problem has been to relate historical costs to future prices. Both problems have multiplied the amount of testimony and greatly prolonged cross-examination and rebuttal hearings.

The costs of the proceedings have been very extensive. The Federal Power Commission's own expenses on gas regulation in the last few years have included more than $3 million per annum incurred for staff devoted to preparing briefs and testimony on area rates as well as for preparation of decisions. The producers have undertaken extremely thorough and expensive investigations of drilling and production expenditures throughout the country. They have retained more than 50 law firms and more than a dozen economic and engineering consulting companies to provide testimony before the Commission. Their costs over all of these area proceedings have never been calculated; but the examiner in the Permian Basin proceedings found that the total costs of producer regulation were 0.14¢ per m.c.f., and uncontested industry testimony in the Texas Gulf Coast Area Rate proceedings showed that these producer costs come to 0.15¢ per m.c.f.[19] The area reviews applied on approximately 75 percent of the 16 trillion cubic feet of producer commitments to the interstate pipelines in each of the last few years, so that total producer expenditures on the area rate reviews must have been close to $18 million.[20]

Since area rates have not been set to apply on all interstate gas sales, producing companies still file some individual rate schedules for review. Also, a number of cases have been filed as exceptions to the Commission-set maximum field prices, for reasons having to do with particular production costs or with specific terms of the sales contracts, or with conditions in state taxation or regulation. These cases have not been dealt with rapidly, presumably because of the priority assigned by the F.P.C. to setting maximum area prices. Consequently, only 500 suspended rate schedules had been dealt with in fiscal year 1968, while more than 7000 other applications for rate increases remained in suspension that year. The inventory of suspended rates has not only been large, but has also involved more than $130 million of disputed payments; the costs of uncertainty and delay for those finally receiving these amounts must be substantial.

---

[19] Cf. *Area Rate Proceeding*, et al. (Permian Basin) Docket No. AR 61-1, 34 FPC 159 (1965) at 197; and the testimony of S. F. Sherwin, Exhibit number 42-J (SFS-1), Schedule 17 in the *Area Rate Proceedings* Docket numbers AR 64-1 and AR 64-2. These average fixed costs of regulation include the costs of certification and of rate schedule review, as well as those in the area rate proceedings, so that they are too large to be an estimate of the private expenses of the rate proceedings alone. But they are not far out of line as a measure of total costs of price regulation because most other schedule proceedings were suspended and replaced by the area rate reviews

[20] The producer commitments to interstate pipelines are given in the F.P.C.'s *Annual Reports.*

At the same time the individual rate reviews have continued to develop into formal producer "rate cases," many involving the documentation of production conditions outside of those setting area rates, or involving company "costs of service" where there are no area rates. Both involve finding "costs" for the *individual* production unit. The arguments for a premium on prices to pay for more costs—say, in more directional drilling for gas in deeper producing regions—have been built on the most extensive statistical and accounting reviews. The documentation inflates the costs of making an individual "rate case" presentation. In the few rate cases for which costs are available, it would appear that defending a rate schedule required expenditures of $25,000; the 80 cases dealt with in a fiscal year result in total defendants' costs of $2 million at least.[21]

The total costs of producer regulation are unknown and can be estimated only very imprecisely. The Commission itself spends $3.1 million per year on all aspects of producer regulation, and the companies probably spend $20 million on defending rate cases—both the large area rate cases and the smaller, more numerous case disputes on suspended individual rate schedules. The costs also incurred by producers from regulating delay at the certification stage are more difficult to estimate; the range given here is from $10 million to $58 million, depending on the length of the delay. As a result, the total regulation costs are from $33 million to $81 million.

☐ **The costs of gas pipeline regulation.** The regulation of the long-distance natural gas pipeline companies resembles that of electric utilities much more than that of the gas producing companies. The market structure in gas transportation and electricity generation are similar. The interstate pipelines have been built to such a scale that some have regional monopolies in supplies to retail gas utility companies, while others face the possibility of substitute sales by no more than two or three other pipeline sources. With few companies in more or less separate sub-national markets, and with each of these companies having some of the characteristics of natural monopolies, the Commission has proceeded with individual "cost of service" regulation.

The pipelines must first obtain "Certificates of Necessity and Convenience" for rights-of-way to construct both their gathering lines in the gas fields and distributing lines to retail gas companies. The hearings and findings of the Commission on these applications are made to determine whether the demands of retail gas utilities are sufficient to require full-capacity operation of the proposed pipeline, and if there are sufficient reserves under long-term field purchase contracts to satisfy these demands. The Commission's quest for quality in service has the effect, of course, of imposing costs on the companies.

---

[21] The source of the data has been private correspondence with a reasonable number of defendants in the cases listed in fiscal year 1968. The costs do not include those incurred in removing rate suspensions in informal Commission proceedings; since more than 500 suspensions are reviewed each year, these total costs could be substantial—albeit unknown at the present time.

In order to prove "sufficiency," the pipeline presents a collection of field purchase contracts and wholesale sales contracts with long lifetimes. It has long been suspected that the time-lengths of these contracts have been longer than would be demanded by the pipelines in the absence of this certification procedure, since the purpose has been to demonstrate full use of the proposed facility for its lifetime rather than least-cost use over the normal decade-long planning horizon in this uncertain industry. The requirement for twenty-year contracts reduces the risks of gas producers, since their sales are secured, and increases the risks of pipeline and retailer not being able to complete final sales of the secured gas at compensatory prices in the far future. The requirement reduces the risks of final consumers of discontinued gas service. These risks will be reflected in costs—higher costs for pipelines and retailers in obtaining capital, and higher prices for the consumer having to pay for these higher costs.

The clerical procedures for proving "sufficiency" are themselves expensive. The contract information is accompanied by geological and legal studies establishing the ownership of gas underground and not yet produced. These studies require the services of experts, and the final reported costs are attributed to the certification process. Part would be undertaken anyway to establish property rights for any gas sale, but the extent of documentation here seems to be much greater. Again, these are costs involved in obtaining more complete knowledge of present and future delivery. There were more than 400 certificate applications before the Commission in 1968 involving 11,000 miles of proposed new line; some were major extensions of more than 100 miles, but most were for a few miles of gathering line into new fields and were not central to the transport of large volumes of gas over 1000-mile main lines.

After certification, the central regulatory concern of the transporters is with the prices to be charged for main-line service. Under the present ratemaking procedures, the company proposes future ceilings on a particular service based on its measures of historical average costs of providing that service. The measures derive from recent expenditures as allocated between types of service according to Commission rules—the most important requiring half of the capital costs to be allocated on the basis of relative volume of gas purchased in the test year, and the other half on the basis of relative volume purchased during peak load periods.[22] The data used in finding the measures, and the actual calculations, are contested by the companies and Commission staff in informal and case rate reviews. Estimates of the rate of return required on pipeline investment by the markets for capital funds depend on assumptions as to alternative uses of funds, and calculations of final "costs of service" depend on how closely the allocation formulas are followed. Both are points of dispute; numerous financial and economic experts describe their judgments on the procedure or estimate those "costs" likely to reduce prices and increase quality in the long run, leaving it to the examiner and Commission to judge the correct values.

---

[22] The so-called "Atlantic Seaboard formula" as given in *Atlantic Seaboard Corp. et al.* 43 PUR (NS) 235 (1952). The costs assigned to service j are

$$C_j = VC_j + K/2[Q_j/(Q_i + Q_j)] + K/2[Q_{jt}/(Q_{it} + Q_{jt})]$$

where VC_j are total variable costs actually incurred in j, k is total capital cost, and Q's are volumes consumed by j and other i demanders throughout the year and also at peak load period t.

TABLE 1

THE ANNUAL COSTS OF REGULATION

ACTIVITY	FEDERAL POWER COMMISSIONS COSTS ESTIMATE (MILLIONS OF $)	COMPANY COSTS (MILLIONS OF $)	
		ESTIMATE	RANGE
ELECTRIC POWER PRICE REGULATION	1.6	1.6	0.2 to 3.1
ELECTRIC POWER SYSTEMS EVALUATION	1.3	1.9	
GAS FIELD PRICE REGULATION	3.1	30.0	30.0 to 78.0
GAS PIPELINE PRICE AND SYSTEMS REGULATION	3.5	2.5	
GENERAL COMMISSION ADMINISTRATION	1.1		
TOTAL ANNUAL COSTS	10 6	36.0	34.6 to 85.5

The Federal Power Commission in recent years has exercised its judgment in a large number of informal rate filing actions as well as formal case reviews of pipeline rates. The natural gas pipeline companies file close to 1500 rate schedules each year, and those dozen or so calling for rate increases are either allowed or suspended by the Commission almost immediately. The suspended rates are either withdrawn or become central issues in the 30 formal rate cases pending each year before the Commission, along with rate decreases called for by F.P.C. staff investigators. When cases are completed, the F.P.C. staff then carries out compliance reviews to ascertain that reductions "flow through" to final consumers of gas.

Altogether, these activities cost the Federal Power Commission close to $3.5 million per annum.[23] The sum total of the private costs of regulation for the gas pipelines was $2.5 million in the last recorded year.[24] Both changes in Commission procedures and in lawyers' and economists' charges can change the total; also, a slight change in the mix towards more complex cases in the range of filings made in a year could revise it upwards by a large amount. But these are not likely events, given that the Commission has worked out highly formal procedures for finding costs and that it entertains differences of views only in narrow topics of cost allocation, tax payments, or in rate of return levels. The 1968 cases were quite orthodox, and they cost the defendants about what can be expected.

There are substantial costs in each of the Federal Power Commission's procedures. A first, but indicative, review of available information on expenditures of both regulator and regulated is shown in Table 1. The costs of government activities are roughly matched by those of private companies in all activities except gas field price regulation, where private expenditures are ten times greater.[25] The costs can vary greatly from those shown; in particular, the range of likely private expenditures is from $35 to $85 million per year. But further analysis is most probably going to find F.P.C. costs slightly lower and private costs slightly greater than shown. The expenditures of all concerned are close to $47 million on Federal Power Commission regulation in a single year.

---

[23] The Federal expenditures for natural gas pipeline regulation are shown as $3.526 million in fiscal year 1968, according to the Budget [22], Appendix, p. 930.

[24] Compiled from "regulatory expense" and consultants' charges shown by each of the interstate pipeline companies in the FPC's "Form 2 Reports" [19] for 1968.

3. The benefits from regulation by the Federal Power Commission

■ The standards for setting maximum prices have been rather general, given the mandates of Congress and the courts, and those for regulating quality have been *ad hoc* by nature. Changes in emphasis on portions of statutory language can lead to significant changes in regulatory procedures and reviews. Given such discretion, what has been the behavior of the Commission? With a well-established organization and peremptory power to demand resources, what has been decided by the Commissioners, and what have been the effects of the decisions on the economy in general?

The Commission itself attempts to show the economic effects of regulation in the *Federal Power Commission Annual Report*. The 1968 *Report* announced that during the last fiscal year "reductions in wholesale rates of electric power moving in interstate commerce and subject to the Commission's jurisdiction amounted to $8,860,595; the largest total for a single year in the Commission's history. . . . While many rate filings were submitted by the utilities on their own initiative, nearly 70 percent of the total dollars of rate decreases were the result of Commission action. . . ."[26] The cases on proposed price increases decided in the Commission's favor, or withdrawn by the companies, resulted in dollar amounts of rate payments that were eliminated rather than incurred by the consumers. These are counted among the gains from power regulation.

There were much greater gains claimed from imposing ceiling prices on field supplies of natural gas. The interim ceiling prices set in 1961 were lower than some unregulated prices on sales of new reserves in 1960. The resulting total "savings" in purchase costs of pipeline buyers were consumer benefits attributed to regulation. The Commission's assessment of the benefits was given in the 1964 *Annual Report* as $434 million,[27] the amount collected by the producing companies subject to refund after the regulatory review was complete.[28]

The Federal Power Commission also saw benefits from reduced gas consumer prices as a result of gas pipeline rate investigations. The Commission states that, "Of ten major proposals to increase [pipeline] rates by a total of $63,380,200 annually . . . one proposal, involving $2,687,100 annually was rejected and two involving $95,400 annually were accepted without suspension. . . ."[29] Another $13 million of rate increases were withdrawn by the companies in the face of Commission opposition, and there were $13.9 million in rate reductions required of the pipelines after Federal Power Commission rate filing reviews. The Commission actions in the gas pipeline industry resulted in $28 million of rate reductions which were eventually to become benefits to the individual consumers of natural gas.

---

[25] The minimum field price costs of $30 million include $10 million resulting from certification delay (assuming that the delay has been cut to a minimum of very few months), $18 million from area rate proceedings, and $2 million in other rate reviews and cases.

[26] From the 1968 *Annual Report* [17], p. 19.

[27] The 1964 *Annual Report* [17], p. 140.

[28] The Commission did not state directly that the receipts subject to refund were consumer benefits. Two implicit statements are needed to make the argument: (1) these receipts $(\Delta P)Q$ for volumes Q at price reduction ΔP are benefits to pipelines, (2) the pipelines pass them on to the consumer in full.

[29] *Annual Report* [17], p. 59.

This accounting results in greater estimated benefits to the economy than economic analysis allows. The rate reductions are realized by some consumers as increases of real income, but other consumers in their roles as stockholders of gas and electricity companies experience income losses from dividend reductions. The income gains of the first group and the losses of the second, shown as the area "A" equal to $(P_1 - P_2)Q_1$ in Figure 1, result in only partial net economy-wide gains. The net gains depend entirely on the extent to which the income distribution after the rate reduction is more "socially acceptable" than the distribution before the reduction.[30] Only the "quantity-quality" increases generated by regulation are net economy-wide gains. The consumption generated by the reduction in price can result in net gains to both groups, given that the area "B" under the demand function is not an income loss of the producers and is a gain of the consumers.[31] The consumption generated by increased demands under regulation, shown as area "C," also result in net gains for the economy. But neither areas "B" nor "C" can be said to be a large part of the Federal Power Commission's dollar savings from regulation.

Not only economic analysis but also conditions in gas and electricity markets may not allow so tolerant an interpretation of the results from regulation. Price reductions might have taken place in the absence of regulation as a result of cost reductions. After all, monopolists are expected to reduce prices when scale effects are predominant, or when factor costs decline as a result of technical progress, or when there is substantial entry of other companies on geographical or product fringes of markets long held on an exclusive basis. All these conditions have been present in these two industries. There is evidence of large scale economies in systems, and of substantial recent cost reductions from fuel price reductions in electricity.[32] Companies in each industry face increased competition from those in the other and from coal and oil in industrial markets for energy. Any company may be more than willing to reduce rates for reasons of profit and credit the results to vigorous regulation by the Commission.

☐ **Economic benefits from electric power regulation.** The Federal Power Commission announced $8.9 million of electricity price reductions in the 1967–1968 fiscal year, but claimed that only 70 percent of this amount resulted from Commission initiatives, so that the benefits from regulation were $6.2 million in that year. This is a first and potentially interesting indication of the size of the price reduction on all established sales (equal to area "A" in Figure 1). But an additional step has to be taken to make it useful in assessing economy-wide gains. These gains of consumers have to be measured against the losses of other consumers receiving income from the electricity-generating companies.

---

[30] There is no *a priori* reason to favor the consumer's income over that of stockholders, because there is no theory establishing the general equity of such redistribution. In fact, there is no theory implying that such redistribution makes incomes more equal, rather, this is an empirical matter.

[31] There are technical assumptions required to make this statement hold, which imply that the effects can be described by movements along this demand curve and not from shifts of the curve. Cf. Friedman [4]. But also the area "B" has to be net of costs of supplying the additional consumption, even though the "costs of service" do not show whether this is the case.

[32] Cf. MacAvoy and Sloss [10], and MacAvoy [8], Appendices A and C.

The calculation proceeds by groups of consumers. The purchasers of electricity that are industrial or commercial enterprises experiencing price reductions surely gain no more than the electricity producers lose. Price reductions are passed on in dividends or lower final-product prices. The recipient companies' stockholders gain the dividends that the electricity company stockholders lose; there is no net gain unless there is some particular condition such as the recipients being "poorer" than the electricity company equity holders. The final consumers of products from companies with lower electricity costs could, of course, capture all the benefits. But this has not been very likely—the four largest gas and electricity consuming industries in the 1960's were so highly concentrated that individual recipient companies certainly need not have passed on small cost reductions in order to survive.[33] It cannot be assumed that an appreciable proportion of the $6.2 million accrues to final consumers through producing companies experiencing lower energy costs; it is expected here in practice that there is no net gain from income transfer on 55 percent of the dollar rate reductions.[34]

The final consumers of electricity, in contrast, may gain some part of the amount that the owners of the electrical companies lose. There are many different opinions on how much a dollar is worth when given to person X after it has been taken from person Y. Here the point of view is that the government should decide, and it indicates value by the amount that consumer X can keep after taxes (the remaining portion of the dollar being taken in taxes for better political use elsewhere).

The governments of this country tax consumers on income received in money wages and salaries, but not income received in price reductions on consumption goods. It would appear that the rate of political preference for price reductions is very high, and any amount passed directly from producer to consumer should be treated as completely gained. The preference for dividend increases, as revealed by tax treatment of these incomes, is very low; dividends are taxed at the corporate rate of 50 percent before they are distributed,[35] and then a second time at the applicable personal income tax rate. It would appear that only the income left after double taxation is lost by a rate reduction.

As a result, reducing dividend incomes of electricity company stockholders by $1.00 transfers $[0.50(1 - t)]$ ($1.00) of "politically approved" income to consumers.[36] With an average tax rate "t" on dividend receivers' own personal incomes equal to 40 percent,[37] the transfer to consumers results in gains of 30¢ on the dollar. The other 70¢ on the dollar is received by the consumer, but it comes from taxes foregone and the electricity consumer has no special claim on that over recipients in government programs. The amended estimate of net gains on direct consumer sales in the \$6.2 million is \$0.8 million.[38]

---

[33] According to the Census Bureau [14], the four largest consuming industries were primary metals, chemicals, petroleum, and stone-clay-glass. The four digit concentration ratios in these industries were 56.1, 49.3, 36.7, and 37.1 percent, respectively, according to the Census Bureau [15], they are all higher than the all-industries average of 32 percent.

[34] In a typical year in the late 1960's, commercial and industrial sales accounted for 55 percent of total revenues from ultimate or final service each year. Cf. [20].

[35] Only so long as companies cannot take advantage of exceptions such as depletion or depreciation allowances. These exceptions, of course, reduce the real rate below the formal 50-percent rate.

[36] That is, the direct transfer from net income of dividend receiver to net income of rate payer is $1.00 minus taxes, or $(\$1.00) - [(1.00)0.50 + (1.00)(0.50)t] = 0.50(1 - t)(1.00)$.

The economic gains are increased, however, by the quantity increases generated by the rate reductions (shown as area "*B*" in Figure 1). The maximum estimate of this surplus is the area $(\triangle P \cdot \triangle Q)/2$ for the regulation-induced price reduction $\triangle P$ and the resulting increase $\triangle Q$ in quantity demanded. The Commission reports do not estimate the rate reductions, nor their effects on quantity demanded. They can be roughly approximated in percentage terms.

The Commission tells us that rate reductions in the two most important cases were 5 percent and 9 percent of total receipts, so that the average of $(\triangle P)(Q)/(PQ)$ can be assumed to be 7 percent. Other sources provide indications of demand elasticity $P\triangle Q/Q\triangle P$. Demand elasticity estimates in detailed econometric studies range from -0.2 in the high-income Northeast states to -0.7 in the Southern states for home consumption. Industrial demand studies indicate elasticities of -1.0 in ten large using industries.[39] A less exact but more current overall estimate of elasticity is -1.2.[40] The 7 percent and the -1.2 can be used to find area "*B*." Multiplying this last elasticity estimate by the percentage reduction provides an estimate of ΔQ; this along with the Commission's announced reductions of \$6.2 million provides an estimate of $(\Delta P)(\Delta Q)/2$ of \$260,000.[41] The quantity-increase benefits from one year's electricity regulation are not much more than one quarter million dollars.

The total annual benefits in "*A*" and "*B*" should last at least five years, and it may well be ten years before changes in market structure and demand conditions vitiate the one year's particular events in regulatory history. Accounting for this lag effect, the total benefits from the surveillance and rate review activities during a single year are equal at most to the present value of ten years of price reduction. An example is shown in Figure 2; at time t_0, a review of rates has been completed and has resulted in prices being reduced from P_1 to P_2; both P_1 and P_2 fall after that time as a result of greater than economy-wide productivity increases, but P_2 falls slower than P_1 until at time t_1 much of the effect of the rate review has been eroded; at that time a new review takes place. The example raises the central questions on how rapidly the two price lines converge and when the next review takes place. The answers can only be surmised at this

---

[37] The average income of all stockholders is estimated to be \$108,000 per year. This income is from weighting incomes shown in *Statistics of Income* [16] for 1967 (Individual Returns) by the amount of dividends received. The average aggregate tax rate on this income, given personal deductions allowed in a four-person family, cannot exceed 40 percent. Then the implied rate of taxation on \$1.00 of pre-tax dividend income of an electric generating company is 50 percent on corporate tax and 20 percent on personal tax (or 40 percent on the remaining after-corporate-tax 50 percent).

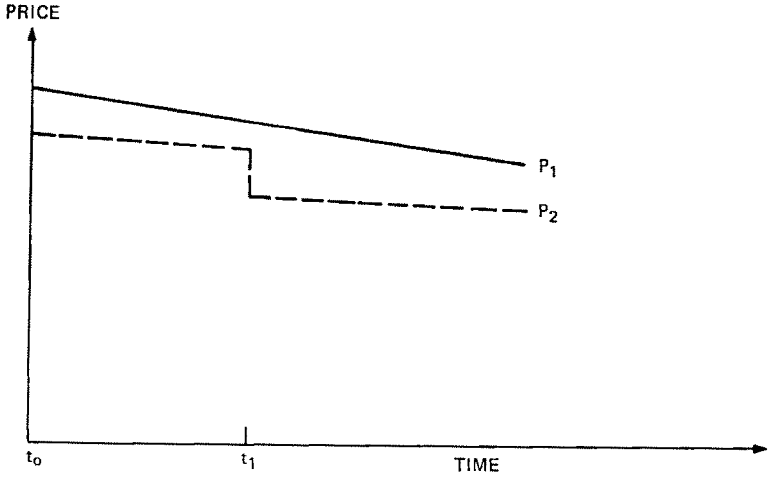
[38] This is 30 percent of the 45 percent of total sales accounted for by home consumers.

[39] See Fisher and Kaysen [3], pp 42, 50, 135. These are averages over states and industries, roughly in accordance with the amount of electricity consumed.

[40] See MacAvoy [8], Appendix C, "Documentation of Forecasts." The elasticity is for price of electricity effects on the demand for generating capacity, and this is a long-run elasticity.

[41] The calculation procedure is straightforward. There are three pieces of information. (a), $(\Delta P)(Q) = -(6.2)(10^6)$, (b), $(\Delta P)Q/PQ = -0.07$; and (c), $P\Delta Q = -1.2Q\Delta P$ assuming that the elasticity of final demand is -1.2. From (b) and (c), $-P\Delta Q/PQ = -0.084$ or $Q = \Delta Q/0.084$. Substituting this expression for Q in (a), then $(\Delta P)(\Delta Q) = -(6.2)(10^6)(0.084)$ and ½ of this amount constitutes a first estimate of consumers' surplus.

FIGURE 2



stage of our knowledge of F.P.C. activities; a conservative response—making for a maximum estimate of benefits—is that it takes ten years for all effects of a review to be realized, but the effects are discounted at 15 percent per year.[42] The present value of income benefits of type "*A*" is $4.0 million, and of quantity type "*B*" is $1.3 million.

These amounts must be maximum net benefits. The price reductions required after any one Commission review are usually small, and the costs of litigation not so small, so that the temptation of the company is to concede the Commission-ordered reduction even when it takes some rates below costs.

This is seen in the most recent formal case review of an electric power company's rate schedule before the Commission: the Northern States Power Company (in Docket E-7140) conceded $254,468 of rate reductions based on 1963 schedules after an extensive cost analysis by the company showed "cost of service" of $4.653 million and a parallel analysis by the Commission staff showed "costs" of $3.486 million. There were no standards by which the difference could be resolved, since they "reflect differences in cost assignment and allocation procedures" (34 FPC 883, at 884). Without substantive procedures, the company could only assess its chances of winning in Court in a random process but after extensive additional costs of litigation. It chose to concede "without prejudice to its contentions" and the Federal Power Commission found this "an acceptable compromise of contested issues involving numerous judgmental factors" (*ibid*). The company's choice was between the costs of litigation, or revenue losses in price reduction, whichever were less, even though the second resulted in rates below production costs. The clear possibility exists that these reductions took rates below production costs, with consequent loss of some part of the quantity gains in area "*B*."

There are two more reasons why this is a maximum estimate of the gains from electricity rate regulation. First, some part of the "Commission-induced" reductions would have occurred anyway. One company president said, in response to requests for estimates of the private costs of regulation, "in the early 1960's the Federal

---

[42] The 15-percent discount rate is somewhat high, but deliberately so. There are two reasons. this is a realistic estimate of the rate of return in the private sector on resources drawn off for the regulatory process, and this accounts for the possibility of very high rates of decay in the gap between P_1 and P_2 when the economy changes rapidly.

Power Commission for the first time asserted jurisdiction over the sales by electric utility companies to local distribution systems. . . . Although the Commission's jurisdiction had been defined by statute in 1935, the Commission under Chairmen Swidler and White greatly extended the concept of what constituted an interstate sale. In most states these wholesale rates had been regulated by the state commissions for many years. . . . Many state commissions still assert duplicate jurisdiction over such sales. . . ."[43] With duplicate regulation, some rate reductions occur as a result of the activities of state commissions.

Last of all, quite plausible conditions in electricity markets implying lower benefits could hold rather than those assumed above. The percentage price reduction, and the elasticity of consumers' demand with respect to that reduction, could well have been much lower than the estimates used here. In fact, price reductions below five percent, with a value of elasticity of -0.2 (as found by Kaysen and Fisher for home consumers in high-income states) make the estimated consumers' surplus "*B*" from regulation so close to zero that it might as well not be considered at all. A more cautious view of benefits restricts the estimated amount to that for area "*A*," for a five-year period before the "decay" of the difference between regulated and unregulated prices, and this amounts to only $2.7 million from one year's regulation.

☐ **Economic benefits from gas field price regulation.** The Federal Power Commission's setting of interim natural gas field prices was not enough of an "act of regulation" by itself to show economic benefits. Since these interim prices are still being reviewed in the "area rate proceedings," regulatory effects have not yet been realized. But timing alone is not the only reason for lack of results; in fact, regulatory gains are not going to be realized in any straightforward manner whenever the area rates are determined. The effects of area rates include hard-to-detect gains by some consumers and losses by other consumers as well as producers.

The proposed area rates freeze prices at the level attained on the larger "packages" of field reserves in the late 1950's to very early 1960's. Those prices "cleared the market," given the demand and supply conditions of that time. Since then market conditions have not been the same—population and income increases, changes in tastes in favor of cleaner fuels, have increased demands by substantial magnitudes. The question is whether there has been additional supply forthcoming to satisfy the greater demands for new contract reserves. Under normal conditions of gas discovery, with effective competition in field markets,[44] the imposition of 1960 price ceilings could only guarantee that the quantity supplied would reflect 1960 conditions. Then 1970 (increased) demand and 1960 supply would imply excess demand. But gas discovery and development vary greatly from year to year without close and direct relation to prices, so that supply could have increased more or less than normal.

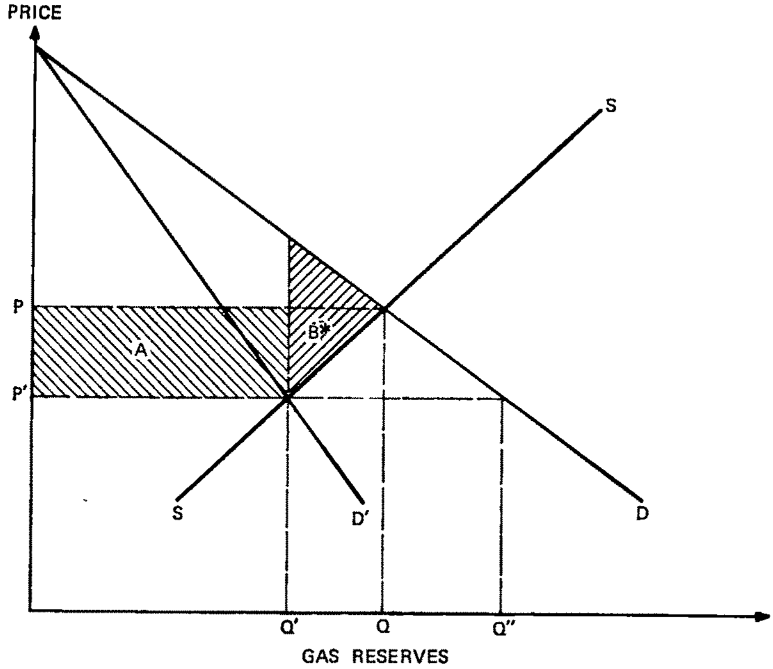
Excess demand where it occurs results in reduced benefits to consumers who must do without some part of the service they seek at

---

[43] Private correspondence initiated by the author

[44] The effectiveness of competition here is one of the most researched topics in industrial economics in the last few years, there is tacit agreement even in the area rate cases to center analytical attention on the competitive model. Cf. review and citations in Kitch [6].

FIGURE 3



the going prices. The lower regulated prices provide gains for those established consumers not affected by service restriction. But the lower prices also reduce incomes of those receiving dividends from the gas-producing companies. The gains have to be compared with the losses.

These gains and losses are illustrated in Figure 3 for a regional market containing new reserves of gas for sale under long-term contract to pipelines. The demand for reserves of gas to serve new industrial and home consumers is shown as the curve *D*, with additional reserves sought by pipelines at lower prices per thousand cubic feet. The supply of these reserves—found in exploration for gas and oil and developed by additional drilling beyond the exploratory well—is shown by the curve marked SS. The market clears at price *P* cents per thousand cubic feet of reserves and the quantity *Q* trillion cubic feet of reserves dedicated to pipelines in new contracts in a single year. The Commission, however, sets area rates that average only *P′* so that the quantity demanded is *Q″* while the quantity supplied is *Q′*.

Consider first the possibility that each consumer receives that amount of gas he would have gotten in an unregulated auction of amount *Q′*. With these restricted supplies, the smallest possible loss to excluded consumers is that shaded area *B** in the diagram,[45] since this is the graphical representation of the net amount over costs that all of the (unsatisfied) consumers would have paid to receive the "lost" output *Q* − *Q′*.

---

[45] Area *B** is not equal to area *B* in Figure 1 because it includes both triangles above and below the price line *P*. The triangle below the price line, equal to producer's surplus, was excluded in the first diagram. This was not for analytical reasons but because there is no measure at this point in research for the difference between marginal costs (in that case) and "costs of service" (equal to average price in that case). At the least, the companies have not announced their gains from the quantity or sales increases initiated by regulation (if any).

Now consider the possibility that some consumers receive as much as they wish while others go without. The favored consumers have demand D'. The greatest possible loss is the whole area between the two curves D and D' above the supply curve; this would occur if new groups of consumers are excluded outright and all of the demands of other consumers are met.[46]

Either of the loss estimates have to be balanced against the potential gains to those receiving all they want. These gains are that "politically acceptable" part of the income transfer shown as area A; they are only part because some of the transfer is to other industries besides the gas producing industry, and some is foregone tax payments given to consumers with "socially-less-desirable" incomes. We shall attempt to estimate the losses from excess demand and this income gain as well.

The losses of unsatisfied gas consumers in recent years have not been directly observable. Even if they were great, the losers themselves have not been aware of their circumstances because the pipelines have been providing them gas that they will not be able to provide in the future. Faced with new contract volumes less than the amounts necessary to make deliveries to new customers over the next twenty years, the pipelines provide gas to all new sources of demand for a shorter time period. The losses to customers are from the shortened commitment—from gas not available to continue delivery in the eighteenth, nineteenth, and twentieth years. Most of them do not react to these losses when delivery begins; but they can still be detected indirectly in the supply-demand behavior of the gas field markets.

Here the losses are assessed by first finding (1) the magnitude of excess demand $(Q'' - Q')$, then estimating (2) the area under the demand function D between Q' and Q'', and last subtracting out (3) both the area under the supply curve SS between Q' and Q and the area under the demand function D between Q and Q''. This procedure leaves the shaded area B^* in Figure 3 as a residual. So as not to exaggerate the losses, this residual is made as small as seems reasonable.

1 *Excess demand* $(Q'' - Q')$ depends on the amount of new reserves needed to meet commitments to new buyers and on the amount needed to replace depleted reserves for established buyers. The Federal Power Commission considers reserve demand of pipelines to be twenty times the volume of initial delivery to a final buyer; that is, demands for reserves are based on "the assumption that each new market commitment is backed by a twenty-year supply."[47] To make this commitment on new sales to industrial and retail utility consumers, new reserves should be purchased by the pipelines equal to twenty times the amount of any additional delivery to final consumers; this "commitment to expanded delivery" is derived as demand $D(1)$ in column (3) of Table 2 by multiplying year-to-year production changes by twenty. To maintain commitments to established buyers, the same rule would apply on replacement reserves. Each year, approximately 5 percent of deliveries into pipelines deplete old reserve commitments of the gas producers, and

---

[46] This is the case in which pipeline extensions to new locations are simply not allowed because "reserves are insufficient."

[47] See [21], p. 18.

TABLE 2

ESTIMATES OF SUPPLY AND DEMAND FOR RESERVES IN THE UNITED STATES

YEAR	SUPPLY	DEMAND	
	ADDITIONS TO RESERVES (TRILLIONS OF CUBIC FEET)	D (1) RESERVES REQUIRED FOR ADDITIONAL PRODUCTION (TRILLIONS OF CUBIC FEET)	D (2) RESERVES REQUIRED TO REPLACE DEPLETED 1962 RESERVES (TRILLIONS OF CUBIC FEET)
(1)	(2)	(3)	(4)
1962	19.5	18.2	13.6
1963	18.2	16.0	13.6
1964	20.2	18.2	13.6
1965	21.3	25.8	13.6
1966	20.2	17.8	13 6
1967	21.8	19.9	13.6
1968	13.6		

NOTE· COLUMN (3) CONSISTS OF ADDITIONS TO PRODUCTION IN THE YEAR FOLLOWING THAT YEAR LISTED IN COLUMN (1), MULTIPLIED BY 20. FOR EXAMPLE, THE NET CHANGE IN NATIONAL PRODUCTION OF GAS IN 1963 WAS 909 BILLION CUBIC FEET, SO THAT REQUIRED RESERVES IN 1962 EQUAL (909) (20) OR 18 2 TRILLION CUBIC FEET.

SOURCE: AMERICAN GAS ASSOCIATION RESERVES OF CRUDE OIL AND NATURAL GAS IN THE UNITED STATES, VOLUME 23, MAY 1968.

the continuation of service to final consumers requires their replacement with new commitments of 20-year reserves with the same annual rate of production. Demand $D(2)$ for commitments to these final consumers equals 5 percent of "base year" (1962) deliveries times twenty, or is equal to the actual levels of delivery in that year.[48] These demands are shown in column (4) of Table 2. The two demands together have been for more reserves than those likely to have been supplied from newly developed gas acreage by an amount of 13 trillion cubic feet per year in the period 1962–1967.[49]

This amount can be called "excess demand" only if two conditions hold. First, there should have been no excess holding of reserves from earlier years to be put against the deficit in the 1962–1967 period. Second, there should have been continued demands of established buyers for the gas listed in "replacement reserves" even when there were substantial price increases. The first would seem to have been the case; prices were generally increasing in the five years previous,[50] and there were deficits in reserves in each of these years averaging to 4.0 trillion cubic feet per annum.[51] But the second condition may not have been realized. The price increases experienced in the 1950–1960 period should have been sufficient to induce many of the gas users in the primary metals industries to switch to residual fuels when long-term contracts under the old prices expired; indeed, the volume of use of gas in that industry declined from 3.8 trillion cubic feet in 1954 to 0.9 trillion cubic feet in 1962.[52] This was extra-marginal replacement, and there were other industrial buyers in disadvantageous locations or with low-priced alternative fuel suppliers who were in the same category.[53] If most of replacement demand actually did disappear at prices close to the F.P.C.'s ceiling prices, then there was no excess demand. As one alternative plausible state of the industry, it is assumed here that most of the replacement demand shown in Table 2 existed, and thus that excess demand has been 13 trillion cubic feet per annum. As another alternative, less plausible state of this industry, it is assumed that there was very little replacement demand and thus that there was no excess demand.

---

[48] If production on old contracts equals X, an amount that is 5 percent of committed reserves R in every case, then new reserves to maintain the rule-of-thumb should be $\Delta R = 0.05R = X$.

[49] The average difference between (a) reserves in column (2) and (b) "commitments" in column (3) plus column (4), was 12.7 trillion cubic feet per year

[50] Cf. MacAvoy [9], Chapters 5, 6, and 7.

[51] The 1955–61 total deficit of "additions to reserves" under "demand for reserves" is estimated as 28.1 trillion cubic feet as calculated from American Gas Association statistics, as noted in Table 2.

2 *The area under the demand function* over the range of excess demand can be approximated if estimates are available of the amount of excess demand (13 trillion cubic feet to zero trillion cubic feet), and the amount of "cleared" demand (approximately 19 trillion cubic feet of new reserves on average as shown in column (2) of Table 2). Also, there has to be an estimate of the elasticity of demand for additional reserves.[54] In fact, there are many demand studies of final users of gas, the most complete and analytically convincing being those of Pietro Balestra. [2] His studies of incremental demands of home consumers provide an indication of the demands for new reserves to be used to provide more home consumption; they show a price elasticity of -1.3 in the last year of the study (1962).[55] Demands of industrial users, accounting for almost 65 percent of the total volume of production at the present time,[56] have not been analyzed in a dynamic model; but studies of industrial buyers at different locations with varying gas prices show price elasticities in consumption exceeding -2.7.[57] The weighted average of home and industrial elasticities is close to -2.2.[58] Given this estimate, the gross loss of consumers' surplus $\triangle P \triangle Q/2$ from excess demand is estimated at $\triangle P \triangle Q/2 = P'(\triangle Q)^2/2Q''e_D = \$204$ million.[59]

3 *The net loss is equal to this gross loss on sales to consumers not receiving service, minus the costs that would have been incurred on the additional service, and minus the demand at prices below the market clearing level.* To make calculations of these areas, all that are needed, in addition to the estimates above, are the elasticity of the supply function e_s and the market clearing quantity Q at the unregulated price. To find the elasticity of supply, we turn again to the F.P.C.

---

[52] Cf. [14].

[53] In fact, the all-industries consumption of natural gas shown in the Census of Manufactures [14] declined from 5.9 trillion cubic feet in 1954 to 4.3 trillion cubic feet in 1962. This partial census alone implies that (1.6 trillion) (20) = 32 trillion cubic feet of demand in that period was eliminated.

[54] The area under the demand curve between Q' and Q'', and above the existing regulated price P', is approximated by $\Delta P \Delta Q/2$. Here $\Delta Q = Q'' - Q'$, and $\Delta P = P' \Delta Q/Q e_D$, with e_D equal to the elasticity of demand for new reserves. Then $\Delta P \Delta Q/2 = P'(\Delta Q)^2/2Q''e_D$.

[55] Balestra [2], pp. 95–99.

[56] The F.P.C. *Annual Report* [17], 1968, p. 45

[57] Cf. the F P.C.'s econometric model, as in *Testimony of* J. Harvey Edmonston, Federal Power Commission Docket AR 61-2, "South Louisiana Area Rate Proceeding." The model contained in this testimony has been severely criticized because of feedback from demand to supply that always "clears the market" at the demand-determined price But the industrial demand sub-system has not been open to such criticism, indeed, it is used here because it is the same as many other studies of industrial demand for gas

[58] The weighted average follows from assuming that household and commercial demand both have the same elasticity and comprise 35% of the total new demand for reserves The industrial demand includes electric power and gas transmission demand and comprises 65% of total consumption Consumption and reserve demand are assumed to be the same in all cases—an assumption that holds only if there is no decline in industrial consumption with rising prices, as above. See the Commission's 1968 *Annual Report* [17], p. 45.

[59] Here P' the regulated price is 17 cents per m.c.f., ΔQ or excess demand is 13 trillion cubic feet, Q'' the total of realized and excess demand is 32 trillion cubic feet, and e_D is -2.2.

econometric model of gas prices. There we find that reserves $\{R = 3.5 + 0.257f \cdot Y\}$ where f is footage drilled per well and Y is the number of discovery wells, a factor dependent upon both price and production; removing the defective feedback from production (demand) to reserves (supply) in Y results in a direct price-supply relationship[60] with an elasticity close to $+ 0.239$.[61]

The market clearing quantity Q can be found indirectly. The change in price P to bring forth supply is equal to the change in price required to reduce demand to the market clearing level, so that $\Delta P/P = (Q - Q')/Q' \cdot 1/e_s = -(Q'' - Q)/Q'' \cdot 1/e_D$. This equation can be solved for Q. For the indicators available here, the estimated market clearing quantity Q is close to 20 trillion cubic feet in a "typical" year in the late 1960's—an amount very close to the 19 trillion cubic feet found on average under regulation.

All the estimates are now complete for measuring the "shaded area" of losses to unsatisfied consumers in Figure 3. The area under the supply curve is approximately $18 million, a small amount reflecting the low elasticity of demand[62]; the area of excess demand below price P is $174 million, reflecting the high elasticity of demand.[63] The losses to consumers from excess demand are $204–192 million, or $12 million per year. This is the greatest likely dollar equivalent to the shaded area in Figure 3. The smallest likely loss is zero, as a result of approximate equality of Q', Q, and Q'' from regulated prices clearing the market at the approximate unregulated level.

No one knows how long such annual losses will continue to be part of the regulatory results in setting gas field prices. At one extreme, they could continue for the lifetime of an area rate schedule set this year, with the schedule lifetime extending to ten years. At the other extreme, 1970 would be the last year of excess demand if the Commission moved rapidly to raise the interim rate ceilings in the pending cases and to revise upwards those in the Permian Basin decision. The area rate proceedings themselves have taken or will take more than five years to complete, so that the costs of litigation in "one year's regulation" have been repeated five times over; for comparability of one year's costs of regulation with benefits from that one year's activities, the one to ten years of benefits have to be divided by five. The negative benefits for unsatisfied consumers range from zero to $12 million.[64]

---

[60] Cf. the testimony of P. H. Cootner in rebuttal of J. Harvey Edmonston, Docket AR 61-2, "South Louisiana Area Rate Proceeding." Professor Cootner recalculates the equation after removing this feedback condition, and the recalculation is used here.

[61] In shortened form $Y = \alpha P^{\beta}$, where α stands for a series of variables unrelated to price, and $R = 3.5 + 0.257(\alpha P^{\beta}) f$. Then $e_s = P\partial R/R\partial P = \beta(0.257\alpha f P^{\beta})/R = \beta(R - 3.5)/R$. In this case, taking $R = Q' = 19(10^{12})$ and $\beta = 0.293$ (from P. H. Cootner, *op. cit.*) results in $e_s = + 0.239$.

[62] This area is found by solving $PQ/2 = P'(Q' - Q)2Q'e_s$ with the values in the text.

[63] This area is equal to $PQ/2 = P'(Q'' - Q)^2/2Q''e_D$ for unregulated price P', the total of realized and excess demand Q'' and market clearing demand without regulation Q. This area is not consumers' loss. It is excess demand that would not be satisfied in an unregulated market because costs are greater than unregulated prices for these amounts.

[64] The difference between this calculation problem and that for benefits from electricity regulation follows from the fact that the area rate cases are not complete, while those analyzed for benefits from electricity regulation have been completed. The assumption leading to the maximum loss of $12 million is that the area rate reviews will be complete after five years, and that ten years of present value of negative benefits will follow. Then "one year's" benefits from regulation are

$$\frac{1}{5} \sum_{1}^{10} B_t/(1 + r)^t$$

for 10 years at $r = 15\%$. In this calculation this equals $B_1(1.0037)$ where $B_1 = \$12$ million.

The income gains of established consumers. Some consumers have been able to obtain the amounts of gas they demanded, and they have gained real income from the price reductions imposed by the Federal Power Commission. Their benefits are not likely to be equal to the amounts now collected by the producers "subject to refund" after the area rate proceedings are complete, because these amounts imply much lower ceiling prices than those likely to be put into effect. In fact, the ceiling prices set in the Permian Basin area rate proceeding were equal to those on non-regulated transactions in the early 1960's; as Commissioner O'Connor states in his separate opinion in that proceeding: "From the record in this proceeding, for the period with which we are concerned, the overwhelming volumes of gas well gas are sold under new contracts at the weighted average price of 16.5 cents (the area rate)."[65] As a result, most "refundable collections" on Permian Basin production will not be refunded, and similar findings on the ceiling rates elsewhere will have similar results.

The consumers' income gains have been made on contracts signed for production of reserves after 1965. The amount of the gains depends on the prices that would have been set by purchasers in the absence of regulation—on price P rather than P' in Figure 2. If the circumstances were those of no excess demand, the gains were zero because prices P and P' were approximately the same. If the circumstances were those of excess demand in keeping with $12 million losses to unsatisfied consumers, the market price P would have to have been some two cents higher than the regulated price. The difference $(P - P')$ on the completed sales Q' was the maximum income redistribution gain for "satisfied" consumers.

The consumers' gains here were producers' or government losses. Since the case can be made for net gains only on sales to home consumers, they were limited to that part of new reserve sales made to interstate pipelines for resale to home consumers.[66] In each of the last few years, approximately 18 trillion cubic feet of new reserves were dedicated to interstate pipelines under the rate ceilings, and perhaps 35 percent of this amount was for delivery to home consumers.[67] Then the gains were limited to two cents per thousand cubic feet on 6.3 trillion cubic feet, or $126 million each year.[68] The losses in government tax receipts were 50 percent of this price reduction, because the petroleum companies would have paid income taxes in excess of the depletion allowances on this amount as if it had been their income and then the dividend receivers would have paid income taxes on it once again. The net income redistribution gains—reallocating only the non-tax losses of dividend receivers—have been limited to $63 million at a maximum (if there has been a gas shortage) and zero at a minimum (no shortage).

---

[65] *Area Rate Proceeding* No. AR 61-1, 34 F.P.C. 159, at 265. Similar findings apply to "old" gas production.

[66] The argument here is the same as for income transfers from lower electricity prices all industrial producers or consumers are in the same income class, but home consumers are in a lower income class; the value of transfer to the lower income class is shown by the marginal tax treatment of a dollar there rather than to the gas company's dividend recipients.

[67] This percentage assumes that residential and commercial sales are all "home sales." Cf. statistics on gas reserves and sales for resale to home consumers in the 1966, 67, and 68 *Annual Reports* [17].

[68] As in the case of income transfers from reduced electricity prices, it is assumed that consumers gain $(0.50 - t)$ dollars for every dollar of reduced dividend income, and that $t = 0.20$ is the foregone tax rate on the dividend. Here it has been assumed that $(0.80 - t)$ holds for depletion revenues, and $t = 0.30$ because of higher incomes earned from the lower company taxes.

The problem with these income gains is that they have been realized on the basis of only interim ceiling prices, and these prices may be lower than the final area prices set in the formal rate proceedings. Additional rate reviews are likely to result in smaller price reductions in face of pressures from excess demand. The chance of higher prices is great enough that future gains must not only be discounted at the 15-percent rate but must be assumed to be limited to ten years. Also, five years are required to complete any of these proceedings, so that the total gains for established consumers have present value of approximately \$63 million.[69]

The net benefits to consumers as a group could have been equal to the gains of the advantaged of \$63 million minus the losses of the deprived of \$12 million. At least such an estimate is plausible, if not convincing. As likely is the possibility that there was no excess demand, so that gains and losses were both zero.

☐ **The economic benefits from natural gas pipeline regulation.** The Federal Power Commission's proceedings on applications by the interstate gas pipelines had the effect of reducing some prices and limiting entry in some instances. The amounts of the price reductions may have been greater than those associated with regulation of electric power prices or gas field prices. The savings from prohibiting entry are unknown.

The Federal Power Commission disallowed \$2.7 million of price increases proposed during the 1968 fiscal year, as well as \$12.0 million of previously proposed increases.[70] They were responsible in some part for \$13.9 million of rate reductions the last year and \$50.9 million the year before, "which were, in large part, flowthrough of supplier refunds and rate reductions."[71]

All of these reductions might be credited to the Commission. None of them might have taken place in the absence of regulatory price reviews, even though it could have been profitable for the firms to reduce prices by the announced amounts given cost reductions from increased advantages of scale or other aspects of technical progress. But even given such a view, credit can be taken in pipeline regulation only for the first and second reductions, since the third and fourth already have been attributed to field price regulation; and this credit is for deliveries at 10% under unregulated prices at the most, so that the net gains from more consumption $\Delta P \Delta Q/2$ could have equalled no more than \$2 million.[72] The accompanying gains from distributing stockholders' income to consumers were limited by the small consumer participation in final demand; with 35 percent of final demand attributable to home consumers and 30 percent of that equal to the indirect income gains from redistribution to "poor" consumers from "rich" stockholders, the net benefits in one year are estimated at

---

[69] This calculation assumes that it will take five years of regulatory expenditure to attain ten years of gains of \$63 million per annum. Both gains and costs occur over roughly the same time period—although the gains occur mostly at a later date. As a first approximation, then, costs for one year of regulation times five equal total regulatory expense, and one year's benefits of \$63 million times ten equal total benefits. For convenience, neither costs nor benefits are given in total value terms here, but rather in terms of one year's results.

[70] See the 1968 *Annual Report* [17], p. 59.

[71] *Ibid.*, p. 60.

[72] This calculation proceeds in exactly the same fashion as those for finding consumers' surplus from electricity rate reductions. Given that $(\Delta P)(Q) = \$14.69$ million, $\Delta P/P = 10$ percent, and the elasticity of demand is -1.28 as shown by the econometric demand relations for gas described above, then $\Delta P \Delta Q/2 = (-0.10P)(-1.28QP/P) = (-0.10)(-1.28\Delta P \cdot Q) = \$1.88$ million.

$1.5 million.[73] At most, ten years of gains could result from one year's regulation, so that the present value of the year's activities in setting rates for natural gas pipeline companies would be $9.4 million for "quantity-increasing" regulation and $7.5 million from "income redistributing" regulation.

The gains from regulating entry are another matter. Without any regulation at all in the pipeline industry, it would be expected that local monopoly pipelines would make extra-normal profits until other, newer lines impinged on their market areas. The additional transporters would reduce profits to normal levels by reducing prices, but also by duplicating the facilities of existing transporters to the point where average mileage costs move up to the existing price level. But with effective price regulation, there should be strong disincentives for such entry to take place, because regulated prices would be reduced to the level of lowest costs for the single transporter. There is no need for strict entry controls designed to prevent duplication unless price regulation is ineffective.

The "price reducing" and "quality" regulation activities of the Commission have been mutually exclusive, particularly because so little has been done except rate and entry limitation in the gas pipeline industry. Both have had the goal of reducing costs from competition when there have been real gains possible from economies of scale in the single transporter. In the precedent cases, the proceedings have unfortunately centered on limiting competition by making judicial choices as to who should be the single source of supply, but without reference to who has the lowest expected economic costs. Russell has shown this in three important and recent certification case decisions of the Commission.[74] The main issue was not encountered: to the extent that price regulation was effective in these instances of two transporters seeking to provide the same service, entry regulation was redundant.

The Federal Power Commission makes no claim for cost savings from limiting entry. Also, the Commission reports no significant improvements in the quality of service from more regulation; indeed, the impetus for quality improvement in pipeline safety has come not from the regulatory Commission but from Congress and private researchers. All that can be assumed at this point is that the gains have all been made by price regulation, whether or not price control has been "perfect."

A review of costs and benefits, and an economic view of commission reform

■ The impression gained from the Federal Power Commission's activities, and the responses of companies in the electric power and natural gas industries, has been one of vigorous and meticulous enforcement of regulation. All that activity cost at least $46 million during a typical year in the late 1960's. The expenditures were made to protect the consumer, and some consumers appear to have gained while other consumers and stockholders lost. The benefits from Commission-initiated price reductions were not only diverse within a Commission program, but also less than the costs of regulation in some but not in other programs.

---

[73] That is, $(0.35)(0.30)(\$14.69)(10^6)$ equals $1.5 million.

[74] Cf. Russell [12], pp. 33–38.

TABLE 3

THE ANNUAL COSTS AND BENEFITS OF REGULATION

FEDERAL POWER COMMISSION ACTIVITY	F.P C. COSTS ($ MILLIONS)	COMPANY COSTS ($ MILLIONS)	BENEFITS ($ MILLIONS)	
			INCOME REDISTRIBUTION	EFFICIENCY
1. ELECTRICITY PRICE REGULATION	1 6	1.6 (0.2 to 3.1)*	4.0 (2.7 to 4.0)*	1.3 (0 to 1.3)*
2. ELECTRIC POWER SYSTEMS EVALUATION	1 3	1.9	* * *	* * *
3. GAS FIELD PRICE REGULATION	3.1	30.0 (30.0 to 78.0)*	63.0 (63 0 to 0 0)	–12.0 (–12.0 to 0.0)
4. GAS PIPELINE PRICE AND SYSTEMS REGULATION	3 5	2 5	7 5	9.4
5. GENERAL COMMISSION ADMINISTRATIVE SERVICES	1 1	0.0	* * *	* * *
TOTAL	10 6	36 0 (34.6 to 85.5)*	74 5	– 1.3
ALTERNATIVE TOTAL (ASSUMING ZERO BENEFITS ON NUMBER 3.)			11.5	10.7

* RANGE OF COSTS OR BENEFITS *** NO ESTIMATE AVAILABLE

The detailed accounting for benefits and costs in Table 3 points out the areas of Commission activity most out of balance. The Commission's activities in the pricing of electric power required $3 million of litigation expenses, half of which were outlays of the regulated companies defending themselves in "cost of service" reviews or rate cases. In return, the consumers received $4 million of indirect additions to income and $1 million of surplus from (imputed) gains in output. It would appear that $3 million of legal resources were used to produce only $1 million of additional net value of product here. The rest of the gains were from F.P.C. taxation—from reducing the incomes of dividend recipients and increasing those of consumers. Other tax authorities have the same mandate, and most probably lower costs for completing the mandate.[75]

Field price regulation procedures have been burdensome in every case, beginning with certification of firm sales of new reserves and continuing with the rate proceedings. The Commisison spent more than $3 million on these activities in 1968. The indirect expenses of the companies, from delays in production imposed by certification and review, and the direct expenses of certification requirements and litigation, probably came to $30 million that year. The benefits directly depended on the extent of price reductions achieved under area rate rulings. If prices were so reduced as to create shortages, then consumer losses on output exceeded $12 million and income distribution gains exceeded $63 million. Once again, extensive resources (almost $33 million) were used up to redistribute amounts of income not much greater ($51 million). But there may have been no excess demand from reduced prices; in that case there were no benefits (as shown by the alternative totals in Table 3).

---

[75] The costs of particular procedures in power regulation are perhaps much greater than benefits from those particular procedures, and much greater than the average benefits and costs show. Correspondence with the power companies on their costs of regulation pointed to certification procedures which duplicated the work of state commissions completely with zero benefits.

In comparison, the results from pipeline regulation were excellent. For no more than $6 million of annual outlay on certification and rate review, the Commission made contributions to consumer rate reductions modestly estimated to be worth more than $9 million in increased consumers' surplus and $7 million in income redistribution.

The Federal Power Commission must make its own calculations, if only to provide a check against those attempted here. If they do, they may again see the effects from the imbalance of regulatory activities. Gas field price regulatory activities must be ten times more costly than those in the "orthodox" public utilities; that alone should raise questions as to whether the bureaucracy either in or engendered by field price regulation has not grown too fast and too large. The costs of redistributing income must be more than $33 million—far more than any levy for collecting comparable incomes by the Internal Revenue Service.

The question is whether the Commission and the rest of the Government see in the imbalance of benefits and costs from gas field regulation a lesson in economic reform. Those activities which make the direct and indirect expenses of producer regulation so great could be curtailed. The case for price and quantity review on the signing of a sales contract by a (regulated) producer with a (regulated) pipeline buyer is exceptionally weak. A social costs-receipts analysis of the crudest sort shows the unsatisfactory state of affairs here. This puts the alternative to the Commission: regulation of field prices by automatically certifying all competitive transactions, so that other agencies then can move on to redistributing income. The means are available for doing so in Commission interpretation of the Supreme Court's mandate for producer regulation; as elucidated by Commission O'Connor in the Permian Basin Case, "a *market* price has been established which provides a reasonable measure of the rate necessary to elicit supply. . . . There is no substantial difference between market and costs in the Permian. Either method is presently permissible for purposes of area ratemaking."[76] The Federal Power Commission has the power to reduce its own scale of activities by centering its factual inquiry on determining the competitiveness of markets and then approving all competitive market prices. The costs of regulation would be reduced to a size comparable to those in electric power and gas pipeline regulation. Thus the policy question has to do with the wisdom and courage required of an agency to reduce itself to one fifth present scale.

The same "agonizing reappraisal" should be conducted on electricity price regulation. This regulation applies only on interstate transmission of power for wholesale distribution to local utilities, so that the market conditions and the issues are not quite the same as those in finding ceiling prices on thousands of gas field sales for the first time. But the greater part of the benefits are from income redistribution, and these taxation-like gains are at the expense of more than $3 million of litigation and administrative resources.

There are two arguments, and as many counterarguments, for continuing the day-to-day surveillance procedures in electricity regulation. First, ongoing regulatory activities have the effect of preventing widespread evasion of the rules given in the important precedent cases. But suppose widespread evasion existed? This deterioration of regulation could lead to prices perhaps 15 to 20 percent higher on wholesale distribution of power before it would cost more *in resource misallocation* than the present costs for *litigation and rate surveillance.*[77] Second, the procedures resulted in the consumer obtaining more than $4 million of present value of price reductions in a typical year. But if this were not the case, then the government itself would have had the present value of twice this amount in tax receipts on "extra normal" electricity profits in a typical year. Admittedly, these taxes could still be collected. But this is a two-sided argument: the "tax" collected by the F.P.C. and passed through to the consumers could also have been collected by the Internal Revenue Service via direct profits levies, and with lower collection costs. The arguments for F.P.C. procedures are, at the most, unsettled in the case of wholesale electricity regulation.

---

[76] Commissioner O'Connor, *op. cit*, p. 265 (emphasis added).

Then what can the Commission do? The mandate is present in this social accounting for continued day-to-day control of pipeline rates and service. The mandate is present and unused for improving the quality and adding to the quantity of service provided by the pipeline and electric power companies. The next decade will see the Federal Power Commission engaged in the search for more low-cost power and gas—where the costs include those imposed on others using air and water resources appropriated by the utilities. Otherwise, the next decade will see the Commission not at all.

References

1. AMERICAN GAS ASSOCIATION *Reserves of Crude Oil and Natural Gas in the United States.* Vol. 23, 1968.
2. BALESTRA, P. *The Demand for Natural Gas in the United States.* Amsterdam: North Holland Publishing Company, 1967.
3. FISHER, F. M. and KAYSEN, C. *A Study in Econometrics: the Demand for Electricity in the United States.* Amsterdam: North Holland Publishing Company, 1962.
4. FRIEDMAN, M. "The Marshallian Demand Curve," *Journal of Political Economy,* Vol. 57, No. 6 (December 1949), pp. 463–95.
5. GERWIG, R. W. "Natural Gas Production: A Study of Costs of Regulation," *The Journal of Law and Economics,* Vol. 5 (October 1962), pp. 69–92.
6. KITCH, E. W. "Regulation of the field Market for Natural Gas by the Federal Power Commission," *The Journal of Law and Economics,* Vol. 11 (October 1968), pp. 243–80.
7. LIVERSIDGE, A. "Not Enough Gas in the Pipelines," *Fortune,* Vol. 80, No. 6 (November 1969), p. 120.
8. MACAVOY, P. W. *Economic Strategy for Developing Nuclear Breeder Reactors.* Cambridge, Mass.: The M.I.T. Press, 1969.
9. ———. *Price Formation in Natural Gas Fields.* New Haven, Conn.: Yale University Press, 1962.
10. ——— and SLOSS, J. *Regulation of Transport Innovation: Unit Trainloads of Coal to the Eastern Seaboard.* New York: Random House, 1966.
11. MAIN, J. "A Peak Load of Trouble for the Utilities," *Fortune,* Vol. 80, No. 6 (November 1969), p. 116.
12. RUSSELL, M. "Resource Allocation in Utility Certification Decisions," *M.S.U. Public Utilities Studies* (East Lansing, Michigan: Michigan State University Press, 1969).
13. SCHMALENSEE, R. L. "Regulation and the Durability of Goods," *Bell Journal of Economics and Management Science,* Vol. 1, No. 1 (Spring 1970), pp. 54–64.

---

[77] The vertical distance or ordinate in triangle "*B*" in Figure 1, when the area of that triangle does not exceed the costs of regulation of $3 million.

14. U. S. BUREAU OF THE CENSUS. *Census of Manufactures, 1963.* Washington, D. C.: U. S. Government Printing Office, 1966.
15. ———. *Concentration Ratios in Manufacturing Industry, 1963.* Washington, D. C.: U. S. Government Printing Office, 1966.
16. U. S. DEPARTMENT OF THE TREASURY. *Statistics of Income.* Washington, D. C.: U. S. Government Printing Office, appropriate years.
17. U. S. FEDERAL POWER COMMISSION. *Annual Report* Washington, D. C.: U. S. Government Printing Office, annual.
18. ———. *Electrical Utilities and Licensees (Classes A and B), Annual Report.* Washington, D. C.: U. S. Government Printing Office, annual.
19. ———. *Natural Gas Companies (Classes A and B)* Washington, D. C.: U. S. Government Printing Office, annual.
20. ———. *Statistics of Electric Utilities* Washington, D. C.: U. S. Government Printing Office, appropriate years.
21. ———. BUREAU OF NATURAL GAS. *A Staff Report on Natural Gas Supply and Demand.* Washington, D. C.: U. S. Government Printing Office, September 1969.
22. U. S. GOVERNMENT PRINTING OFFICE. *The Budget of the United States Goverment, 1970.* Washington, D. C.: U. S. Government Printing Office, 1970.
23. WINCH, D. M. "Consumers' Surplus and the Compensation Principle." *American Economic Review*, Vol. 55 (1965), pp. 395–423.

A mathematical formulation of the peak-load pricing problem

Israel Pressman

Assistant Professor, Polytechnic Institute of Brooklyn
and
Management Sciences Division
American Telephone and Telegraph Company

This paper formulates peak-load pricing problems using mathematical micromodels. The optimal strategy chosen for the public utility is that of maximizing the social satisfaction derived from services provided. The notion of consumers' surplus is used, and period demands are assumed to be both independent and dependent. The case of dependent demands, a heretofore unsolved problem, is handled using the line-integral calculus. Several specific models are analyzed, with both capacity constraints and profit constraints being considered. In some models it is shown that prices should depend on marginal operating costs but not on marginal capacity costs. Trade-offs between these marginal costs are explored in the peak-load pricing context. In other models the relationships of price with both demand elasticities and marginal costs are developed. Several of the existing peak-load pricing models of the literature can be shown to be subcases of models developed herein.

1. Introduction

■ The demand for public utility service varies not only periodically over the month and year, but also varies significantly on a daily or weekly basis. Public utilities have been concerned for nearly a century with determining methodologies, techniques, and philosophies for coping with this peaking of demand. The problem of meeting these variations in load with some optimum sized plant capacity and the accompanying investments and costs, all in the framework of a pricing structure, is called the peak-load pricing problem.

The maximum capacity of the plant and the pricing of the various utility services are highly dependent on the set of assumptions used to define the peak-load pricing problem. Suppose for example, that a single service is to be produced during two periods for which only two costs are incurred. Let b be the operating cost per unit per period and let β be the cost of providing a unit of capacity.[1] For the moment, both b and β are assumed to be constant. It has been shown that in the long run the marginal cost of a unit of output will be b (which is the short-run marginal cost) if capacity is not fully utilized and $b + \beta$ (which is the SRMC plus the marginal cost of a unit of capacity) if all the existing capacity is used up.[2] If the demand curves

---

Dr Pressman holds the B S. and M.S degrees in Physics and received the Ph D. degree in Operations Research from The Johns Hopkins University in 1968. He has taught Physics at Brooklyn and Queens Colleges, and has been a systems analyst at the IIT Research Institute and the RCA Service Co His present interests include computed-assisted systems for teaching and medical diagnosis and economic modeling of the regulated public utility.

The author is indebted to the editors and to several colleagues who commented on drafts of this paper, but particularly to Mr Steven M. Fulda of the A T. &T Co., who provided both the definition of the problem and guidance throughout the research

[1] See section 6, model, D.

[2] Steiner [22], p 587.

are given for each period, it is possible to determine the optimal output in each period and the prices associated with these outputs. The capacity required will be the peak demand on the system. Using these particular assumptions, if only for illustrative purposes, the peak-load problem will exist at any price if the demanded outputs in the two periods at that price are different.

In general, however, several difficulties arise when more realistic assumptions are used. One of the difficulties is the choice of a proper and meaningful operating criterion for the firm. Others include choosing proper constraints, defining the cost function, and providing an economic interpretation of the results.

One of the most important assumptions that must be closely examined is the dependence-independence property of the demand functions. In much of the literature of mathematical economics a basic assumption is that the demand for a good or service depends only on that good or service. In reality, however, demand functions are not independent but depend on all the goods and/or services demanded in the multi-period problem involved. Thus, for a regulated telephone utility that sells a single service in two distinct periods known as the peak and off-peak periods, each demand function would depend on both the peak and off-peak demand.

Several articles have appeared [1], [6], [23], [26] which examine and solve parts of the peak-load pricing problem. Boiteux [1] establishes a framework for using marginal-cost pricing for the peak-load pricing problem. Steiner [22] solves the case for linear costs with a capacity constraint. He presents the results graphically. Williamson [26] discusses the welfare motivation of the analysis and extends the linear costs case to handle periods of unequal length. De Salvia [6] demonstrates the applicability of peak-load pricing concepts to a typical U. S. electric utility.

The intent of this paper is to examine the peak-load pricing problem analytically, in a general form, taking into account the special demand properties and the possible constraints that are necessary for a meaningful framework for the problem.

2. The objective function

■ As pointed out previously, one of the difficulties in a peak-load pricing problem is choosing the "right" operating criterion, which is usually referred to as the objective function. From among the list of possible objective functions that might be used, this study chose finding an optimum level for value of service as its only objective function. This choice was made in light of recent research by public utilities on the value-of-service concept and its quantification and the historical evidence of the usefulness of value of service in defining the public utilities' goals.[3] This paper presents its own interpretation and quantification of value of service to be used in the objective function.

Some of the problems with using this objective function have developed because of difficulties in determining an appropriate, meaningful and workable definition for value of service. The concept itself goes back to the early days of public utility pricing policy and rate making. It has long been recognized both by utilities and commissions that certain public utility services are more valuable to some types of customers than to others. In the communication business, for example, this value of service has developed along the following lines: Based on the socioeconomic usefulness of a telephone to a business customer, basic telephone service is considered more valuable by the businessman than by the typical residence telephone user. Over decades, business telephone rates have developed to be somewhat above residence rates.

---

[3] For example, see F. C. C. Docket 16258.

Thus, the different customers demonstrate the value of service to themselves by their usage of service at the prices charged. This value-of-service concept recognizes, through prices, the relative values that customers assign to each public utility service, and many utilities have long relied on this concept for developing differential prices among their various service offerings.

It is, unfortunately, difficult to quantify value of service, as described above, in strict analytic terms. Some economists have attempted to quantify value of service using elasticity of demand as a rationale for monopoly pricing policy. One definition of value of service that lends itself more readily to reasonable quantification uses the economic notion of the consumer's surplus concept, that is, the difference between what a consumer would be willing to pay for a good or service and what he actually pays. Thus, from the consumer's point of view, the consumer's surplus is a measure of the net benefit he derives from buying a certain quantity of good or service. If we consider the benefit to both the consumer and the producer, then the total gross benefit will be defined as the consumer's surplus plus the total revenue (total revenue equals the producer's surplus plus the total cost). The net benefit will be defined as the gross benefit minus the total cost. Hence, for those who desire to use the value-of-service objective as the goal of the firm, one objective function that is both meaningful and quantifiable is

$$\max(\text{net benefit})$$

which is equivalent to

$$\max(\text{gross benefit} - \text{total cost})$$

which is equivalent to

$$\max(\text{consumer's surplus} + \text{total revenue} - \text{total cost})\,.$$

At this point, before proceeding with the discussion of the various models, it is appropriate to discuss the mathematics and economics of the consumer's surplus concept. It will be shown that the utilization of consumer's surplus in analytical modeling depends on special conditions with a special economic interpretation.

3. Mathematical formulation of the consumer's surplus

■ In the previous section we defined consumer's surplus as the difference between what a consumer would be willing to pay for a good or service and what he actually pays. In actuality, this definition is one of several found in the literature[4] and depends on the economic conditions assumed. An acceptable mathematical formulation for this definition can be found by examining Figure 1, which depicts a demand curve in the price-quantity plane. Let P^* be the

---

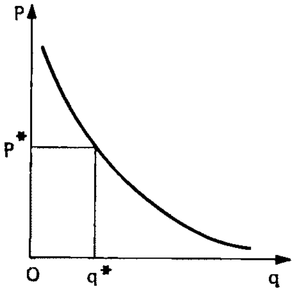
[4] See Samuelson [20] p 197, Hicks [11], p. 40, Pfouts [18] and Winch [27].

equilibrium price the consumer actually pays for the equilibrium quantity q^*. For each quantity q, $0 < q < q^*$, the consumer would be willing to pay an excess price $P(q) > P^*$ rather than go without it entirely. Thus by buying q^* units at price P^* the consumer accumulates some satisfaction in the form of excess price. This satisfaction has been designated the consumer's surplus. In our formulation, following the Marshallian definition rather than the compensating variation concept of Hicks, this surplus is thus equal to the area under the demand curve above the price line. Denoting consumer's surplus by S, this becomes

$$\begin{aligned} S &= \int_0^{q^*} [P(q) - P^*]dq \\ &= \left\{\int_0^{q^*} P(q)dq\right\} - P^*q^* . \end{aligned} \tag{1}$$

Thus, the consumer's surplus is an economic measure of the difference between the gross benefit of the money expended and the total revenue, P^*q^*, from the sale of q^* units.

FIGURE 1



Although other mathematical formulations of the consumer's surplus have appeared,[5] they all (including equation (1)) suffer from one basic deficiency. This deficiency is the assumption that the demand for the ith good or service depends only on the quantity of that good or service, and *not* on any of the other n goods or services. With this assumption, the consumer's surplus for each good or service can be calculated separately and is given by

$$S_i = \int_0^{q_i^*} P_i(q_i)dq_i - P_i^*q_i^* . \tag{2}$$

The total consumer's surplus is then the sum of all the individual surpluses and is given by

$$S = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} \int_0^{q_i^*} P_i(q_i)dq_i - \sum_{i=1}^{n} P_i^*q_i^* . \tag{3}$$

Steiner[6] attempts to extend the consumer's surplus concept to interdependent demands by defining the gross benefit for each good or service as

$$F_i = \int_{q_i} P_i dq_i \tag{4}$$

where

$$P_i = P_i(q_1, q_2, \ldots, q_i, \ldots, q_n) . \tag{5}$$

The total gross benefit would then be

$$\sum_{i=1}^{n} F_i = \sum_{i=1}^{n} \int_{q_i} P_i dq_i . \tag{6}$$

This, however, is only a simple extension of the first term on the right-hand side of equation (3) with P_i redefined as in equation (5). It will be shown in the appendix that this representation, although mathematically convenient, is inadequate for a reasonable economic interpretation of the multi-good consumer's surplus.

---

[5] See Marshall [15], Hicks [10], and Samuelson [20].

[6] In [22], p. 608.

A more convenient formulation for the multi-dimensional consumer's surplus with dependent demand functions is given by Hotelling [14] as a line integral of the form

$$S = \int_C (P_1 dq_1 + P_2 dq_2 + \cdots + P_n dq_n) - \sum_{i=1}^{n} P_i^* q_i^* \qquad (7)$$

with the integration being performed along some curve C and where P_i is defined by equation (5). Thus, for example, if $P_1(q_1,q_2)$ and $P_2(q_1,q_2)$ are two demand functions of q_1 and q_2, then the total consumer's surplus between $(q_1,q_2) = (0,0)$ and $(q_1,q_2) = (q_1^*,q_2^*)$ is

$$S = \oint_{0,0}^{q_1^*,q_2^*} [P_1(q_1,q_2)dq_1 + P_2(q_1,q_2)dq_2] - (P_1^*q_1^* + P_2^*q_2^*) \quad (8)$$

with C some designated path between these two points.

There are two difficulties with this formulation. First, the necessary conditions for a local maximum or minimum of a real function require that the first derivatives of that function vanish at all critical points. Since S is a part of the objective function which is to be optimized, the differentiability of S is required if classical optimization techniques are to be employed. Second, S, as it now is defined, depends on the particular path C chosen and is thus not unique.

Both difficulties are solved by introducing a theorem on line integrals from integral calculus. (See the appendix for a detailed exposition.) The theorem states that if we assume a set of integrability conditions, i.e.,

$$\frac{\partial P_i}{\partial q_j} = \frac{\partial P_j}{\partial q_i}, \qquad (9)$$

then line integrals of the form (7) are differentiable and their value S is independent of the path C. Thus, in order to adapt to our problem the Hotelling formulation of consumer's surplus, the integrability conditions (9) will be assumed. In the next section the economic rationale and interpretation of these conditions is given.

4. Economic rationale for integrability conditions

■ Having shown that a useable formulation for the gross surplus exists only if the integrability conditions are satisfied, it is necessary to examine what economic phenomena would produce these conditions. The economic rationale that follows is due to Hotelling.[7]

Suppose there are N competing buyers of n commodities and that the αth buyer, by purchasing $q_{1\alpha}, q_{2\alpha}, \ldots, q_{n\alpha}$ of the various commodities, can produce and sell goods that will bring him a gross money income u_α. If we write the gross income as a function of the $q_{i\alpha}$'s, then the net income for the αth buyer is

$$\pi_\alpha = u_\alpha(q_{1\alpha}, \ldots, q_{n\alpha}) - P_1 q_{1\alpha} - \cdots - P_n q_{n\alpha}. \qquad (10)$$

Each buyer attempts to maximize his net income by purchasing $q_{1\alpha}, \ldots, q_{n\alpha}$ at prices $P_1, \ldots, P_n$, which he regards as fixed by the competition and not affected by his purchases. By differentiating

---

[7] See [13], p. 67.

(10) with respect to $q_{i\alpha}$ we get

$$\frac{\partial \pi_\alpha}{\partial q_{i\alpha}} = \frac{\partial u_\alpha}{\partial q_{i\alpha}} - P_i = 0\,.$$

If we define $f_{i\alpha}$ as

$$f_{i\alpha}(q_{1\alpha}, \ldots, q_{n\alpha}) = \frac{\partial u_\alpha}{\partial q_{i\alpha}}, \tag{11}$$

then we have as the demand functions for the αth buyer

$$P_i = f_{i\alpha}(q_{i\alpha}, \ldots, q_{n\alpha})\,. \tag{12}$$

From (11) and (12) it is evident that

$$\frac{\partial P_i}{\partial q_{j\alpha}} = \frac{\partial P_j}{\partial q_{i\alpha}}. \tag{13}$$

Equation (12) can also be solved to yield

$$q_{i\alpha} = F_{i\alpha}(P_1, P_2, \ldots, P_n)\,. \tag{14}$$

Differentiating (12) with respect to P_k yields n equations in n unknowns

$$\sum_j \frac{\partial P_i}{\partial q_{j\alpha}} \frac{\partial q_{j\alpha}}{\partial P_k} = \delta_{ik}\,. \tag{15}$$

Solving (15) for $\frac{\partial q_{j\alpha}}{\partial P_k}$ we find that because of (13)

$$\frac{\partial q_{j\alpha}}{\partial P_k} = \frac{\partial q_{k\alpha}}{\partial P_j}, \tag{16}$$

which are the integrability conditions for (14). Let

$$q_i = F_i(P_1, P_2, \ldots ,P_n), \quad (i = 1 \cdots n) \tag{17}$$

be the total demand functions for the whole aggregate of buyers. Thus, $q_j = \sum_\alpha q_{j\alpha}$. Hence, on summing (16) with respect to α we get

$$\frac{\partial q_j}{\partial P_k} = \frac{\partial q_k}{\partial P_j}. \tag{18}$$

Furthermore, solving (17) for P_i yields

$$P_i = f_i(q_1, q_2, \ldots, q_n), \quad (i = 1 \cdots n) \tag{19}$$

which must satisfy

$$\frac{\partial P_i}{\partial q_j} = \frac{\partial P_j}{\partial q_i} \tag{9}$$

by the same argument used to deduce (16) from (13). We note that the economic arguments leading to these integrability conditions assume that no fixed budget exists that must be exactly met. If this were not the case, then equation (9) would have to be replaced by a looser condition, i.e.,

$$P_i\left(\frac{\partial P_j}{\partial q_k} - \frac{\partial P_k}{\partial q_j}\right) + P_j\left(\frac{\partial P_k}{\partial q_i} - \frac{\partial P_i}{\partial q_k}\right) + P_k\left(\frac{\partial P_i}{\partial q_j} - \frac{\partial P_j}{\partial q_i}\right) = 0\,. \qquad (20)$$

Another way to look at the integrability conditions is to note that (9) will be satisfied if the respective components of the cross effects—the substitution effect and the income effect—are equal. Henderson and Quandt[8] indicate that the cross effects (the change in the demand for one commodity resulting from a change in the price of some other commodity) denoted by $\frac{\partial q_i}{\partial P_j}$ can be separated into two distinct effects, a substitution effect and an income effect. Because of the symmetry of the problem[9] the substitution effect on the *i*th commodity resulting from a change in the *j*th price is the same as the substitution effect on the *j*th commodity resulting from a change in the *i*th price. Since the substitution effects are equal we need only examine the income effects. If the budget is not rigid, then the income effect will be negligible or zero[10] and (9) will be satisfied. If, however, the budget is fixed and rigid, then some of the income effect terms will be significant and (20) rather than (9) will be satisfied.

It is also possible to show that the integrability conditions hold with the assumption of a fixed budget. Hicks[11] has shown this to be valid if the income elasticities, $\frac{y}{q_i}\frac{dq_i}{dy}$, for all goods for the same individual are identical. In addition, it is necessary that either the income distribution remain fixed or that the income elasticities for all individuals be identical.

5. Structure for the analysis

■ In developing the various models, four factors are considered. These are:

1. Objective Function,
2. Demand Function,
3. Cost Function,
4. Constraints.

The objective function is chosen to be identical for all the models and, as given previously, is max(net benefit).

For the demand function we choose to examine both models with dependent demands and models with independent demands. For some of the models these different demand properties do not affect the results because of the assumption made about the integrability conditions.

---

[8] In [8], p. 29.

[9] See [8], pp. 24–29.

[10] The income effect reflects the changes in purchases of q with a change in the price, which causes a change in the real income of the consumer. The real income changes because the budget is fixed and the consumer can now buy more of a commodity whose price has fallen and less of a commodity whose price has risen. However, when the budget is not fixed, then price changes cause no appreciable change in the effective buying power, i.e., the same amount of each commodity can be purchased, thus the income effect will be negligible.

[11] See Hicks [9], Appendix 1.

TABLE 1

A SUMMARY OF THE MODEL STRUCTURE

DEMANDS	COSTS	
	GENERAL	LINEAR
INDEPENDENT	MODEL A (PROFIT CONSTRAINT)	
DEPENDENT	MODEL B (PROFIT CONSTRAINT)	MODEL D (CAPACITY CONSTRAINT)
	MODEL C (PROFIT & CAPACITY CONSTRAINTS)	MODEL E (CAPACITY & PROFIT CONSTRAINTS)

Two types of cost functions are considered in the development of the various models, a general cost function and a more restrictive linear cost function. Restricting our study to the two-period case, with q_1 and q_2 the quantity of a single available service demanded in the first and second periods, the general cost function is composed of two parts, the operating costs and the capacity costs. If K is the maximum capacity of plant in output units, then the structure of the operating cost function for each period is defined by $D_i(q_i, K)$, i.e., the cost function in each period is identical and depends on the amount of the demand q_i and the capacity K. The capacity cost is defined as $\bar{q}(K)$, i.e., the cost of establishing a plant of maximum capacity K. Thus the total cost function for the two-period case is

$$TC = D_1(q_1,K) + D_2(q_2,K) + \bar{q}(K)\,. \qquad (21)$$

If we assume that the operating cost is a linear function of q and the capacity cost is a linear function of K, then the linear cost function is defined as

$$TC = bq_1 + bq_2 + \beta K \qquad (22)$$

where b is the operating cost per unit per period and β is the cost of a unit of capacity.[12]

Three kinds of constraint sets are possible in our model. The first is a capacity constraint of the form

$$q_i \leq K \quad \text{all } i\,. \qquad (23)$$

The second is a profit constraint of the form

$$\pi = \text{total revenue} - \text{total cost} \leq M \qquad (24)$$

where total revenues are $\sum_i P_i q_i$, total cost is given by either equation (21) or (22), and M is a constant representing the maximum allowable profit.[13] The third set contains both capacity and profit constraints of the form (23) and (24). Thus, if we consider all possible combinations of our factors (2 demands, 2 cost functions, 3 constraint sets) we can generate twelve different models. Table I presents a summary of the possible models that can be generated. Each box represents a chosen set of demand-cost assumptions. Since three kinds of constraints sets are possible, each box can contain three models. In the following section the analytical and economic aspects of five of these models are examined in considerable detail.

---

[12] This is the type of cost model that Steiner [22] assumes for his peak-load pricing model.

[13] Equation (24) may be thought of as a regulatory constraint on the profits of the firm, while at the same time recognizing that a firm should be allowed to earn an amount above that determined by social welfare considerations.

6. The models

The general model. Consider the problem

$$\begin{gathered}\max(\text{net benefit})\\ \text{subject to profit} \le M, \text{ and}\\ (\text{demand})_i \le \text{capacity}, \; i = 1,2\end{gathered} \tag{25}$$

where only a two-period case is being considered with the demand in the first period q_1 and the demand in the second period q_2, both non-negative, and with non-negative capacity K. The mathematical representation of equation (25) is

$$\max \int \{ \sum_{i=1}^{2} P_i dq_i \} - C(q_1,q_2,K)$$

$$\text{s.t.} \sum_{i=1}^{2} P_i q_i - C(q_1,q_2,K) \le M \tag{26}$$

$$q_i \le K \quad i = 1,2$$

$$q_i \ge 0; \quad K \ge 0 \quad i = 1,2$$

where $P_i = P_i(q_1,q_2)$ and C is the total cost function. Forming the Lagrangian yields the unconstrained maximization problem

$$\max \int \{ \sum_{i=1}^{2} P_i dq_i \} - C(q_1,q_2,K) + \lambda[\sum_{i=1}^{2} P_i q_i - C(q_1,q_2,K) - M]$$

$$- \sum_{i=1}^{2} \gamma_i (q_i - K) \tag{27}$$

$$q_i \ge 0; \quad K \ge 0 \quad i = 1,2$$

where λ and γ_i are Lagrange multipliers.

If $\int \sum_{i=1}^{2} P_i dq_i$ is concave in q_i, then C has to be at least quasi-convex in q_i and K for the objective function to be concave. In this case, however, the constraint $\sum P_i q_i - C - M \le 0$ is concave in q_i and K and the Kuhn-Tucker conditions for a maximum are necessary but not sufficient. On the other hand, if C is concave (implying increasing returns to scale) in such a way that $\sum P_i q_i - C - M \le 0$ is convex in q_i and K, then the objective function is concave if and only if the gross surplus is greater than the cost for small values of q_i. In either case it is still possible to find the maximum by evaluating the objective function at all critical and boundary points and to use the conditions on the signs of the multipliers to indicate whether the point is a candidate for a maximum or a minimum.

Thus on applying the Kuhn-Tucker conditions for a maximum to (27) we get

$$q_i \ge 0; \quad (1+\lambda)\left(P_i - \frac{\partial C}{\partial q_i}\right) + \lambda \sum_{j=1}^{2} q_j \frac{\partial P_j}{\partial q_i} - \gamma_i \le 0; \quad i = 1,2 \tag{28}$$

$$K \ge 0; \quad (\sum_{i=1}^{2} \gamma_i) - (1+\lambda)\frac{\partial C}{\partial K} \le 0 \tag{29}$$

$$\gamma_i \ge 0; \quad q_i - K \le 0; \quad i = 1,2 \tag{30}$$

$$\lambda \le 0; \quad \sum_{i=1}^{2} P_i q_i - C - M \le 0 \tag{31}$$

For each condition we have equality if the corresponding variable satisfies the inequality and vice versa.

The five models which will be analyzed can be deduced from the general model as follows:

Model A) $\gamma_i = 0,\ i = 1,2;\ \partial P_j/\partial q_i = 0 \quad i \neq j$
Model B) $\gamma_i = 0;\ i = 1,2$
Model C) the general model
Model D) $\lambda = 0;\ \partial C/\partial q_i = b,\ i = 1,2;\ \partial C/\partial K = \beta$
Model E) $\partial C/\partial q_i = b,\ i = 1,2;\ \partial C/\partial K = \beta$

The analysis of each of the models follows.

☐ **Model A. A general—cost independent—demand model.** For this model the total cost, C, has the form of equation (21), the cross-effects are zero, and the capacity constraints are not binding. Several cases must now be examined.

a) $\lambda = 0$. If the Lagrange multiplier is zero (which will occur if the constraint is not active), then equation (26) reduces to an unconstrained maximization problem with the solution, price equals marginal cost, i.e.,

$$P_1(q_1) = \frac{\partial D_1}{\partial q_1}$$
$$P_2(q_2) = \frac{\partial D_2}{\partial q_2} \tag{32}$$

and the marginal cost increase in expanding plant by one unit is offset by a marginal cost decrease in operation costs, i.e.,

$$\frac{\partial \bar{q}}{\partial K} = -\frac{\partial D_1}{\partial K} - \frac{\partial D_2}{\partial K}. \tag{33}$$

The implication of equation (32) is that the rates should be set according to the rules of a purely competitive market at equilibrium, i.e., marginal cost pricing. The literature includes several works discussing the marginal cost pricing concepts.[14] In considering equation (33) recall that the firm has an installed and ready-to-use capacity of K, but is operating at all times below that capacity. Then the marginal cost of adding a unit of this capacity K to active operation will be exactly offset by the marginal cost decrease in total operation costs due to capacity. Thus, although it appears that these capacity marginal costs are free, since the right-hand side of (33) contains negative signs, in actuality it is the total marginal cost due to capacity that is equal to zero.

b) $\lambda = -1$. If the Lagrange multiplier $\lambda = -1$ (which occurs as a solution to equation (29) if the constraint is binding and (33) does not hold) then equation (28) reduces to

$$-q_1\frac{dP_1}{dq_1} = 0 \quad \text{and} \quad -q_2\frac{\partial P_2}{\partial q_2} = 0. \tag{34}$$

This can be written

$$P_i\frac{q_i}{P_i}\frac{dP_i}{dq_i} = 0 \tag{35}$$

---

[14] See Nelson [16], and Turvey [24].

or

$$\frac{P_i}{\epsilon_i} = 0 \tag{36}$$

where ϵ_i is the price elasticity of demand for the ith good or service given by

$$-\frac{P_i}{q_i}\frac{dq_i}{dP_i}.$$

Equation (36) implies that either $P_i = 0$ or $1/\epsilon_i = 0$, i.e., $\epsilon_i = \infty$. To determine when $P_i = 0$ will occur, we re-examine our objective function, equation (27), which reduces, when $\lambda = -1$, to[15]

$$\max \sum_i \left[\int_0^{q_i} P_i(q_i) dq_i - P_i q_i \right] \tag{37}$$

or max(consumer's surplus). From Figure 1 we observe that consumer's surplus is maximum when the price line reaches zero. We note that this occurs independently of the costs and thus yields a welfare gain equal to the entire area under the demand curve. This might be explained by interpreting $\lambda = -1$ as the case when the constraint holds as an identity and is thus not controllable by the optimization procedure. $\epsilon_i = \infty$ will occur if we have perfect competition and the demand curves are horizontal. This implies that the prices have already been set by the market and decisions can't affect the prices. Thus the maximization problem would not be applicable for pricing policy.

c) $\lambda < 0$, $\lambda \neq -1$. If $\lambda < 0$ and not equal to -1, then from equation (28) we get

$$\frac{P_i(q_i) - \dfrac{\partial D_i}{\partial q_i}}{P_i(q_i)} = \frac{\lambda}{1 + \lambda} \frac{1}{\epsilon_i}. \tag{38}$$

From equation (38) the following can be noted: If $-\infty < \lambda < -1$, then since $\epsilon_i > 0$, price will be greater than marginal operating cost. If, however, $-1 < \lambda < 0$, then the right side of (38) becomes negative and price is less than the marginal operating cost.

To obtain more insight on the meaning and range of λ we solve (38) for λ obtaining

$$\frac{\epsilon_i \left(P_i - \dfrac{\partial D_i}{\partial q_i} \right)}{\epsilon_i \left(P_i - \dfrac{\partial D_i}{\partial q_i} \right) - P_i} = -\lambda. \tag{39}$$

Thus $\lambda = 0$ if either ϵ_i is zero or $P_i = \partial D_i / \partial q_i$. In addition $\lambda = -1$ if $\epsilon_i = \infty$ or $P_i = 0$. Finally, we have

$$\lambda = -\infty \quad \text{iff} \quad \epsilon_i \left(P_i - \frac{\partial D_i}{\partial q_i} \right) - P_i = 0. \tag{40}$$

---

[15] The M term in (27) is omitted here since it is a constant and doesn't affect the optimization.

This reduces to

$$P_i - \frac{\partial D_i}{\partial q_i} = \frac{P_i}{\epsilon_i} \tag{41}$$

or

$$\frac{\partial D_i}{\partial q_i} = P_i\left(1 - \frac{1}{\epsilon_i}\right). \tag{42}$$

But

$$MR = P_i\left(1 - \frac{1}{\epsilon_i}\right). \tag{43}$$

Thus

$$\lambda = -\infty \quad \text{iff} \quad MR = \frac{\partial D_i}{\partial q_i}, \tag{44}$$

which is the equilibrium condition for a profit-maximizing firm. Hence, λ appears to correspond to the range of profitability of the firm. For $-1 < \lambda < 0$, profits are positive or negative depending on the particular average cost curve. Thus it is possible to earn positive profits even when price is less than marginal operating cost. If $\lambda = -1$, then profits, M, are fixed, while for $\lambda = -\infty$ the fixed profits, M, are a maximum.

☐ **Model B. A general-cost dependent-demand model.** This model is identical to Model A above with the exception that the demands are dependent, i.e.,

$$\begin{aligned} P_1 &= P_1(q_1, q_2) \\ P_2 &= P_2(q_1, q_2)\,. \end{aligned} \tag{45}$$

and thus the cross-effects are not zero. As in Model A several cases must be examined.

a) $\lambda = 0$. If the Lagrange multiplier is zero, then model B reduces to an unconstrained maximization problem with price equal to marginal cost,[16] i.e.,

$$\begin{aligned} P_1(q_1,q_2) &= \frac{\partial D_1}{\partial q_1}, \\ P_2(q_1,q_2) &= \frac{\partial D_2}{\partial q_2}, \end{aligned} \tag{46}$$

and

$$\frac{\partial q}{\partial K} = -\frac{\partial D_1}{\partial K} - \frac{\partial D_2}{\partial K}. \tag{47}$$

This is the same result obtained in model A for $\lambda = 0$.

b) $\lambda = -1$. If the Lagrange multiplier λ is -1, then equation (28) reduces to

$$\begin{aligned} q_1\frac{\partial P_1}{\partial q_1} &= -\,q_2\frac{\partial P_2}{\partial q_1} \\ q_2\frac{\partial P_2}{\partial q_2} &= -\,q_1\frac{\partial P_1}{\partial q_2} \end{aligned} \tag{48}$$

---

[16] See equations (28) and (29).

or

$$\frac{q_1}{P_1}\frac{\partial P_1}{\partial q_1} = -\frac{q_2}{P_1}\frac{\partial P_2}{\partial q_1}$$
$$\frac{q_2}{P_2}\frac{\partial P_2}{\partial q_2} = -\frac{q_1}{P_2}\frac{\partial P_1}{\partial q_2}. \tag{49}$$

But from (9)

$$\frac{\partial P_i}{\partial q_j} = \frac{\partial P_j}{\partial q_i}. \tag{50}$$

Therefore, equation (49) becomes

$$\frac{q_1}{P_1}\frac{\partial P_1}{\partial q_1} = -\frac{q_2}{P_1}\frac{\partial P_1}{\partial q_2}$$
$$\frac{q_2}{P_2}\frac{\partial P_2}{\partial q_2} = -\frac{q_1}{P_2}\frac{\partial P_2}{\partial q_1} \tag{51}$$

or

$$-\frac{1}{\epsilon_1} = +\frac{1}{\epsilon_{21}}$$
$$-\frac{1}{\epsilon_2} = +\frac{1}{\epsilon_{12}} \tag{52}$$

where $\epsilon_{ij} = -(\partial q_i/\partial P_j)(P_j/q_i)$ is the elasticity of the ith good with respect to the price of the jth good, the so-called cross elasticity of demand.

Several conclusions can be drawn from (52). Since both ϵ_1 and ϵ_2 are positive, ϵ_{21} and ϵ_{12} must be negative. This implies that for the two-good situation the goods can only be substitutes. Now, if we had allowed for three goods q_1, q_2, q_3, then equation (52) would become

$$-\frac{1}{\epsilon_1} = +\frac{1}{\epsilon_{21}} + \frac{1}{\epsilon_{31}}$$
$$-\frac{1}{\epsilon_2} = +\frac{1}{\epsilon_{12}} + \frac{1}{\epsilon_{32}} \tag{53}$$
$$-\frac{1}{\epsilon_3} = +\frac{1}{\epsilon_{13}} + \frac{1}{\epsilon_{23}}$$

for which both substitutes and complements can occur.

The second conclusion we draw is that when $\lambda = -1$ our problem corresponds to the problem of maximum utility with the demand functions homogeneous of degree zero and the income elasticity of demand equal to zero.[17] To see this, consider the demand functions

$$P_1 = P_1(q_1,q_2)$$
$$P_2 = P_2(q_1,q_2) \tag{54}$$

---

[17] We assume zero income elasticity of demand otherwise the demand function would have the form $P = P(q_1, q_2, I)$ and equation (55) would have additional terms.

and apply Euler's theorem for homogeneous functions. We get

$$\begin{aligned} \frac{\partial P_1}{\partial q_1} q_1 + \frac{\partial P_1}{\partial q_2} q_2 &= 0 \\ \frac{\partial P_2}{\partial q_1} q_1 + \frac{\partial P_2}{\partial q_2} q_2 &= 0\,. \end{aligned} \tag{55}$$

Dividing the first equation by P_1 and the second by P_2 and recalling the definition for elasticities we have

$$\begin{aligned} -\frac{1}{\epsilon_1} - \frac{1}{\epsilon_{21}} &= 0 \\ -\frac{1}{\epsilon_2} - \frac{1}{\epsilon_{12}} &= 0 \end{aligned} \tag{56}$$

which is identical to equation (52).

c). $\lambda < 0$, $\lambda \neq -1$. If the Lagrange multiplier is negative but not equal to -1, then equation (28) becomes

$$\left[P_i(q_1,q_2) - \frac{\partial D_i}{\partial q_i} \right](1 + \lambda) = -\lambda \left[q_i \frac{\partial P_i}{\partial q_i} + q_j \frac{\partial P_j}{\partial q_i} \right], \quad i \neq j. \tag{57}$$

Using the definition for elasticity and again noting that $\partial P_j/\partial q_i = \partial P_i/\partial q_j$ from the assumed integrability conditions we get

$$\left(\frac{P_i(q_1,q_2) - \dfrac{\partial D_i}{\partial q_i}}{P_i(q_1,q_2)} \right) = \frac{\lambda}{1 + \lambda} \left[\frac{1}{\epsilon_1} + \frac{1}{\epsilon_{ji}} \right]. \tag{58}$$

In addition, for this case we get

$$-\frac{\partial D_1}{\partial K} - \frac{\partial D_2}{\partial K} = \frac{\partial \bar{q}}{\partial K} \tag{59}$$

and

$$P_1 q_1 + P_2 q_2 - D_1(q_1,K) - D_2(q_2,K) - \bar{q}(K) = M\,. \tag{60}$$

From equation (57) it is clear that if the cross-effects are zero, i.e., demands are independent, we get equation (38). In addition, if K were assumed very large and thus treated as a constant in the optimization process,[18] then equations (38) and (58) could be obtained by solving the problem[19]:

$$\begin{aligned} &\max(\text{net benefit}) \\ &\text{s.t. profit} = M \end{aligned} \tag{61}$$

i.e.,

$$\begin{aligned} &\max \sum_i \int_0^{q_i} P_i(q_i)dq_i - \sum_i D_i(q_i) \\ &\text{s.t.} \sum_i P_i q_i - \sum_i D_i(q_i) = M \end{aligned} \tag{62}$$

---

[18] The effect of the very large K is to assume that $\bar{q}(K) = k$ and $D_i(q_i,K) = D_i(q_i)$. Thus we eliminate $\bar{q}(K)$ from (26) to obtain (62) and (63).

[19] This model, as given mathematically by (62), was presented by Baumol in Exhibit 24, Docket 16258 with the conclusion as given by (38). The Baumol Model is a very special case of Model B.

for independent demands, and

$$\max \int \sum_i P_i dq_i - \sum_i D_i(q_i)$$

$$\text{s.t.} \sum_i P_i q_i - \sum_i D_i(q_i) = M \tag{63}$$

for dependent demands. Formulations (62) and (63) are thus special cases of formulation (26). Equations (38) and (58) are obtained from (62) and (63) by assuming that the Lagrange multiplier, λ, is neither equal to zero nor equal to -1.

The $\lambda \neq 0$ assumption is made since the constraints here are given as equalities. The $\lambda \neq -1$ assumption is made so as to distinguish between an equality constraint and an identity constraint. When $\lambda = -1$ the solution to (62) is given by equation (36)

$$\frac{P_i}{\epsilon_i} = 0\,. \tag{36}$$

But (36) assumes that the constraint, profit $= M$, is an identity, i.e., always true for any set of p's and q's. Since the intent of the constraint was not to create an identity, the $\lambda \neq -1$ assumption is made.

☐ **Model C. Model B with a capacity constraint.** Model C is the most general form of the model and will be analyzed as in the previous models by examining several cases.

a) $\gamma_1 = \gamma_2 = 0$. If the Lagrange multipliers associated with the capacity constraints are zero, then from equation (30) the capacity constraints need not be active. Thus, this model is analyzed as if it contained only the profit constraint and no capacity constraint, i.e., Model B.

b) $\gamma_1 \geq 0,\ \gamma_2 \geq 0,\ \lambda = 0$. If all the multipliers are zero then the results are identical to model B, case *a*.

If either of the γ's are strictly positive, say $\gamma_1 > 0$ and $\gamma_2 = 0$, then equations (28) through (31) yield

$$\gamma_1 = \frac{\partial D_1}{\partial K} + \frac{\partial D_2}{\partial K} + \frac{\partial \bar{q}}{\partial K}\,, \tag{64}$$

$$P_1(q_1,q_2) = \frac{\partial D_1}{\partial q_1} + \gamma_1\,, \tag{65}$$

$$P_2(q_1,q_2) = \frac{\partial D_2}{\partial q_2}\,, \tag{66}$$

$$P_1q_1 + P_2q_2 - D_1(q_1,K) - D_2(q_2,K) - \bar{q}(K) \leq M\,, \tag{67}$$

$$q_1 = K\,; \quad q_2 \leq K\,. \tag{68}$$

From equations (64), (65), and (66) it is clear that the price in period 1 absorbs all the marginal cost due to capacity, while for period 2 the price depends only on the marginal operating cost. The implications of this result on pricing strategy remain to be determined.

If both γ's are strictly positive, then equations (28) through (31) yield

$$\gamma_1 + \gamma_2 = \frac{\partial D_1}{\partial K} + \frac{\partial D_2}{\partial K} + \frac{\partial \bar{q}}{\partial K}, \tag{69}$$

$$P_1(q_1,q_2) = \frac{\partial D_1}{\partial q_1} + \gamma_1, \tag{70}$$

$$P_2(q_1,q_2) = \frac{\partial D_2}{\partial q_2} + \gamma_2, \tag{71}$$

$$P_1 q_1 + P_2 q_2 - D_1(q_1,K) - D_2(q_2,K) - \bar{q}(K) \leq M, \tag{72}$$

$$q_1 = K; \quad q_2 = K. \tag{73}$$

Hence, each period's output is equal to the capacity, and the prices charged in each period are composed of part marginal operating cost and part marginal capacity cost. Since the functional form of D is the same for each period and both q_1 and q_2 are equal to K, (70) and (71) imply that $\partial D_1/\partial q_1 = \partial D_2/\partial q_2$ and thus

$$P_1 - \gamma_1 = P_2 - \gamma_2$$

or

$$P_1 - P_2 = \gamma_1 - \gamma_2 .$$

Hence, the higher priced period would be charged a larger share of the marginal capacity cost than that of the lower priced period. The method of determining which period is higher priced is found by examining the demand requirements, $P(q_i)$, of each period and solving the set of equations (69)–(73). Although the desired demand for each period is different, the actual demand, q_i, satisfied in both periods is the same, i.e., K. Thus charging different prices could, to some extent, be termed differential pricing or "price discrimination."[20]

c) $\lambda = -1, \gamma_1 \geq 0, \gamma_2 \geq 0$. Since part (a) has already considered the case when $\gamma_1 = \gamma_2 = 0$ for all λ, we here consider the case when only one of the γ's is strictly positive. However, from equation (29) since $\lambda = -1$,

$$\gamma_1 = -\gamma_2 \tag{74}$$

which implies that if $\lambda = -1$, $\gamma_1 = \gamma_2 = 0$ is the only solution.

d) $\gamma_1 > 0, \gamma_2 > 0, \lambda < 0, \lambda \neq -1$. For this case equations (28) through (31) yield

$$(1+\lambda)\left[\frac{\partial D_1}{\partial K} + \frac{\partial D_2}{\partial K} + \frac{\partial \bar{q}}{\partial K}\right] = \gamma_1 + \gamma_2, \tag{75}$$

$$P_1 - \frac{\lambda}{1+\lambda}\left[\frac{P_1}{\epsilon_1} + \frac{P_1}{\epsilon_{21}}\right] = \frac{\gamma_1}{1+\lambda} + \frac{\partial D_1}{\partial q_1}, \tag{76}$$

$$P_2 - \frac{\lambda}{1+\lambda}\left[\frac{P_2}{\epsilon_2} + \frac{P_2}{\epsilon_{12}}\right] = \frac{\gamma_2}{1+\lambda} + \frac{\partial D_2}{\partial q_2}, \tag{77}$$

$$P_1 q_1 + P_2 q_2 - D_1(q_1,K) - D_2(q_2,K) - \bar{q}(K) = M, \tag{78}$$

---

[20] The question of price discrimination here has been discussed by both Steiner [22] and Hirshleifer [12]. However, since the actual output in each period is the same, independent of the demand function, this author is suggesting the existence of price discrimination.

$$q_1 = K\,; \quad q_2 = K\,. \tag{79}$$

Multiplying equations (76) and (77) by q_1 and q_2 respectively, adding, and substituting equation (79) we get

$$\begin{aligned}(P_1 + P_2)K - \left[\frac{P_1}{\epsilon_1} + \frac{P_1}{\epsilon_{21}} + \frac{P_2}{\epsilon_2} + \frac{P_2}{\epsilon_{12}}\right]\frac{\lambda}{1+\lambda}K \\ = \left[\frac{(\gamma_1 + \gamma_2)}{(1+\lambda)} + \frac{\partial D_1}{\partial q_1} + \frac{\partial D_2}{\partial q_2}\right]K \\ = \left[\frac{\partial D_1}{\partial q_1} + \frac{\partial D_2}{\partial q_2} + \frac{\partial D_1}{\partial K} + \frac{\partial D_2}{\partial K} + \frac{\partial \bar{q}}{\partial K}\right]K\,. \end{aligned} \tag{80}$$

The first term, $(P_1 + P_2)K$, is the total revenue. The right side of (80) is a total cost if we have constant returns to scale and fixed factor costs.[21] Thus, only if

$$\left[\frac{P_1}{\epsilon_1} + \frac{P_1}{\epsilon_{21}} + \frac{P_2}{\epsilon_2} + \frac{P_2}{\epsilon_{12}}\right]\frac{\lambda}{1+\lambda}K = \text{Const.} \tag{81}$$

will the total revenue either equal the total cost or differ from the total cost by a constant.

The constant in equation (81) will be zero if either $\lambda = 0$ or if equation (52) is valid, i.e.,

$$\frac{1}{\epsilon_1} = -\frac{1}{\epsilon_{21}}\,; \quad \frac{1}{\epsilon_2} = -\frac{1}{\epsilon_{12}}\,. \tag{52}$$

However, the validity of (52) implies that $\lambda = -1$. But the assumption was made that $\lambda \neq -1$, $\lambda < 0$. Thus, (81) is never equal to zero and total revenue will not equal total cost in this instance.

If, however, returns to scale are increasing or decreasing, then the relationship between total cost and total revenue is not clear and a further analysis of the various terms in (80) is required.

☐ **Model D. A linear-cost dependent-demand model with a capacity constraint only.** With the total cost given by equation (22) and the capacity constraint as in equation (23), equations (28) to (31) yield

$$-\beta + \gamma_1 + \gamma_2 = 0\,, \tag{82}$$

$$P_1 - b - \gamma_1 = 0\,, \tag{83}$$

$$P_2 - b - \gamma_2 = 0\,, \tag{84}$$

$$q_1 \leq K\,; \quad q_2 \leq K\,, \tag{85}$$

$$\gamma_1(q_1 - K) = 0\,; \quad \gamma_2(q_2 - K) = 0\,, \tag{86}$$

$$\gamma_1 \geq 0\,; \quad \gamma_2 \geq 0\,. \tag{87}$$

In order to make the problem meaningful, we assume that at least one of the q's is at capacity. This eliminates the case that both γ_1 and γ_2 are zero, a case that would imply that the K in the model was not a variable. Thus, if $\gamma_1 = 0$ and $\gamma_2 > 0$ we get

$$\beta = \gamma_2\,, \tag{88}$$

$$P_1 = b\,, \tag{89}$$

---

[21] Given these two assumptions, it can be shown that marginal cost = average cost and thus $(MC)K = (AC)K = TC$. See Henderson and Quandt [8], p. 66.

$$q_1 < K\,; \quad q_2 = K\,.$$
$$P_2 = b + \gamma_2 = b + \beta\,, \tag{90}$$

Thus period 1 charges a price equal to the marginal operating cost since it does not utilize the full capacity, while period 2, which is restricted in output by the capacity, charges a price equal to the sum of the marginal operating cost and the marginal capacity cost.

If both periods operate at capacity, i.e., $q_1 = q_2 = K$ which implies $\gamma_1 > 0$, $\gamma_2 > 0$, then

$$\beta = \gamma_1 + \gamma_2\,, \tag{91}$$
$$P_1 = b + \gamma_1\,,$$
$$P_2 = b + \gamma_2\,. \tag{92}$$

These equations are comparable to (69), (70) and (71). Thus, each period shares a part of the marginal capacity costs. These solutions have been shown graphically by Steiner [22], but his solutions are not applicable in the more general cases we have discussed in Models A, B, and C.

☐ **Model E. Model D with a profit constraint.** With the addition of the profit constraint, we get,

$$-\beta(1 + \lambda) + \gamma_1 + \gamma_2 = 0\,, \tag{93}$$

$$P_1 - b + \lambda\left[P_1 + q_1\frac{\partial P_1}{\partial q_1} + q_2\frac{\partial P_2}{\partial q_1} - b\right] - \gamma_1 = 0\,, \tag{94}$$

$$P_2 - b + \lambda\left[P_2 + q_2\frac{\partial P_2}{\partial q_2} + q_1\frac{\partial P_1}{\partial q_2} - b\right] - \gamma_2 = 0\,, \tag{95}$$

$$P_1q_1 + P_2q_2 - bq_1 - bq_2 - \beta K \leq M\,, \tag{96}$$

$$q_1 \leq K\,; \quad q_2 \leq K\,, \tag{97}$$

$$\lambda(P_1q_1 + P_2q_2 - bq_1 - bq_2 - \beta K - M) = 0\,, \tag{98}$$

$$\gamma_1(q_1 - K) = 0\,; \quad \gamma_2(q_2 - K) = 0\,, \tag{99}$$

$$\lambda \leq 0\,; \quad \gamma_i \geq 0\,. \tag{100}$$

Several cases will now be examined.

a) $\gamma_1 = \gamma_2 = 0$. If both multipliers associated with the capacity are zero then equations (93) to (100) have meaning only if either of the following is true:

1. $\lambda = -1$
2. $\lambda \neq -1$, $K \gg 0$ and not considered a variable subject to the optimization.[22]

For $\lambda = -1$ we get

$$-\frac{1}{\epsilon_1} = \frac{1}{\epsilon_{21}}\,; \quad -\frac{1}{\epsilon_2} = \frac{1}{\epsilon_{12}}\,, \tag{101}$$

---

[22] By $K \gg 0$ we mean the existence of installed plant of such size that capacity is not reached, and that the firm considers K fixed and therefore outside of the optimization.

which is equivalent to equation (52). For $\lambda \neq -1$, $K \gg 0$, we get,

$$\begin{aligned} P_1 - \frac{\lambda}{1+\lambda} P_1 \left[\frac{1}{\epsilon_1} + \frac{1}{\epsilon_{21}} \right] &= b \\ P_2 - \frac{\lambda}{1+\lambda} P_2 \left[\frac{1}{\epsilon_2} + \frac{1}{\epsilon_{12}} \right] &= b\,, \end{aligned} \tag{102}$$

which is equivalent to equation (58) with

$$\frac{\partial D_i}{\partial q_i} = b\,.$$

b) $\lambda = 0$. If the multiplier for the profit constraint is zero then this model is analyzed exactly as model *D*.

c) $\lambda < 0$. If $\lambda = -1$, the only meaningful case is when $\gamma_1 = \gamma_2 = 0$ which was noted in (*a*) above.

If $\lambda \neq -1$ and $\gamma_i > 0$, then equations (93) to (100) yield

$$\begin{gathered} -\beta(1+\lambda) + \gamma_1 + \gamma_2 = 0\,, \\ P_i - b - \frac{\lambda}{1+\lambda} \left[\frac{P_i}{\epsilon_i} + \frac{P_i}{\epsilon_{ji}} \right] - \frac{\gamma_i}{1+\lambda} = 0 \\ i,j = 1,2\,; \quad i \neq j\,, \\ P_1 q_1 + P_2 q_2 - b(q_1 + q_2) - \beta K = M \\ q_1 = K\,; \quad q_2 = K\,. \end{gathered} \tag{103}$$

From equation (103) we get thus

$$\begin{aligned} &(P_1 + P_2)K - \frac{\lambda}{1+\lambda} \\ &\quad \times \left[\frac{P_1}{\epsilon_1} + \frac{P_1}{\epsilon_{21}} + \frac{P_2}{\epsilon_2} + \frac{P_2}{\epsilon_{12}} \right] K = \left[2b + \frac{\gamma_1 + \gamma_2}{1+\lambda} \right] K \\ &\qquad = (2b + \beta)K\,. \end{aligned} \tag{104}$$

The term $(2b + \beta)$ represents the total marginal cost, i.e., both operating and capacity, from both periods. Thus the interpretations given for equation (80) previously apply here to (104).

7. Conclusion

■ In order to develop models useful for analytical economic analysis of the public utility peak-load pricing problem, we have used an approach which formulates this problem mathematically. As part of this approach we have incorporated several economic and mathematical assumptions necessary for providing a framework for analysis. From among a group of possible objectives we chose an objective or goal for the firm, namely, to maximize social satisfactions from service provided.

To express this objective of the firm mathematically, we chose to utilize the notion of the consumer's surplus from microeconomic theory. Consumer's surplus is defined classically as the difference between what a consumer would be willing to pay for a good or service and what he actually pays. This definition is valid when demands are independent, that is, when the price of each service depends only on the quantity of that particular service, even though other services are being offered concurrently. Since one of the objectives of this paper is to allow for dependent demand functions—i.e., each price depending on the quantities of more than one service—it was necessary to extend the mathematical consumer's surplus concept to multidimensions. The mathematical framework for doing so utilizes line integration. A special set of mathematical-economic conditions, the so-called integrability conditions, were shown to be necessary in order that the line integral be differentiable and the optimization analysis completed. The actual objective in these models is to maximize the net benefit, i.e., maximize the consumer's plus the producers' surplus in a multi-period framework, where demands are different for each period.

Two types of constraints were considered as restricting this maximum net benefit, a profit constraint and a capacity constraint. The profit constraint dictates that the difference between revenues and costs must not exceed a specified level, while the capacity constraint implies that the quantity of output of a service in any period is restricted by the capacity of the plant. Using either or both of these constraints, several peak-load pricing models were generated and studied. In this paper we chose five specific models (as indicated in Table 1 of Section 5) to analyze using Kuhn-Tucker optimization techniques.

In models with only a profit constraint and no capacity constraint the prices depend on the marginal operating costs but not on the marginal capacity costs. In addition, the marginal cost of adding an additional unit to operation appears to be free, but in actuality it is offset by a reduction in the marginal operating cost due to capacity. Furthermore, when the constraint is active the prices will depend on the elasticities of demand ϵ_i when demands are independent, and on both the elasticities and the cross elasticities, ϵ_i and ϵ_{ij} respectively, when the demands are dependent.

In models with only a capacity constraint and no profit constraint, prices are charged according to marginal-cost pricing principles. The peak period is charged the full marginal cost of capacity, while the off-peak period is charged the SRMC or the marginal operating cost. If, however, full capacity is utilized in both periods, then each period shares the marginal capacity cost and the period with the greater actual demand gets charged the larger share.

Finally, for those models with both the capacity and profit constraints in cases where these constraints are binding, the prices depend on both the marginal costs and the elasticities of demand. In addition, when we have constant returns to scale and fixed factor costs the total revenue will not equal total costs and, at best, will differ by a constant. If, however, the returns to scale are increasing or decreasing, the relationship between costs and revenues is unknown. Here one is unable to present a clear economic interpretation of the results, and a further analysis of the various terms in the equations is required.

Appendix

Application of the Line Integral Theorem to the multi-dimensional consumer's surplus function

■ In Section 3 we encountered a difficulty with equation (7); namely, that the integration depends on the path C, and that S cannot be differentiated in this general form. It was thus necessary to impose certain conditions on the S function to make it usable in an objective function. The conditions imposed were the result of applying a theorem on line integrals which is given below.

Independence of Path Theorem for Line Integrals. Let $P_i(q_1, q_2, \ldots, q_n)$ be n functions such that $P_1, P_2, \ldots, P_n, \partial P_i/\partial q_j$ (all i and j) are continuous and single valued at every point of a simply connected region. Then, if and only if

$$\frac{\partial P_i}{\partial q_j} = \frac{\partial P_j}{\partial q_i} \tag{9}$$

will the line integral

$$\int_{a \cdots b}^{q_1^* \cdot q_n^*} (P_1 dq_1 + P_2 dq_2 + \cdots + P_n dq_n)$$

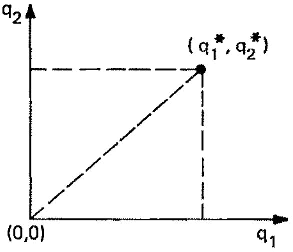
a) be independent of the path from $(a \cdots b)$ to $(q_1^* \cdots q_n^*)$, and
b) be zero around every closed curve in the region.

In addition,

c) there exists a function $F(q_1^* \cdots q_n^*)$ such that $dF = P_1 dq_1 + \cdots + P_n dq_n$ = a total differential and
d) the following equations are satisfied:

$$\begin{gathered}\frac{\partial F}{\partial q_1} = P_1, \ldots, \frac{\partial F}{\partial q_n} = P_n . \\ \text{Note: } \frac{\partial P_1}{\partial q_n} = \frac{\partial^2 F}{\partial q_1 \partial q_n} = \cdots = \frac{\partial^2 F}{\partial q_n \partial q_1} = \frac{\partial P_n}{\partial q_1}\end{gathered} \tag{105}$$

FIGURE 2



This theorem, thus, implies that if conditions (9), called integrability conditions, are satisfied, then we need no longer be concerned with the choice of a path, and that the line integral portion of our S function can indeed be differentiated.

Hence, if the line integral is independent of the path, then the line integral of equation (8) (Section 3) can be though of in two separate ways. Consider the value in moving from point (0, 0) to (q_1^*, q_2^*) on Figure 2 it can be shown that

$$\int_{0,0}^{q_1^*, q_2^*} (P_1 dq_1 + P_2 dq_2) = \int_0^{q_1^*} P_1(q_1, 0) dq_1 + \int_0^{q_2^*} P_2(q_1^*, q_2) dq_2 . \tag{106}$$

Thus, the total value is the sum of two separate values and will be the same regardless of which path is taken between the end points. In addition, this line integral can be thought of as a one-dimensional function along the diagonal where $dF = \overline{P} \cdot d\overline{Q}$ and

$$\overline{P} = \bar{i} P_1 + \bar{j} P_2 \quad \text{and} \quad d\overline{Q} = \bar{i} dq_1 + \bar{j} dq_2 .$$

Thus, the gross surplus would be, mathematically,

$$GS = \int_0^{\bar{Q}'} \bar{P}\cdot d\bar{Q}\,, \qquad (107)$$

an equation similar to what one obtains in the independent case.[23]

The Steiner definition for gross surplus is equation (4) (section 3), and thus we have

$$F_1 + F_2 = \int_0^{q_1^*} P_1(q_1,q_2)dq_1 + \int_0^{q_2^*} P_2(q_1,q_2)dq_2\,. \qquad (108)$$

From Figure 2 this means the surplus is found by integrating along two separate line segments separately and then adding the sum. Since Steiner does not indicate whether the integrability conditions must be satisfied, it is not evident that equation (108) will yield the same result as an integration along some continuous line path. Thus, we believe our choice of the Hotelling formulation for the gross surplus to be the correct one.

References

1. Boiteux, M. "Peak-Load Pricing," in *Marginal Cost Pricing in Practice,* J. R. Nelson (Ed.) Englewood Cliffs, N. J.: Prentice Hall, 1964. pp. 59–90.
2. Buchanan, J. "Peak-Loads and Efficient Pricing: Comments," *Quarterly Journal of Economics,* Vol. 80 (1966), pp. 463–80.
3. ———. "The Optimality of Pure Competition in the Capacity Problem: Comment," *Quarterly Journal of Economics,* Vol. 81 (1967), pp. 703–6.
4. Crew, M. A. "Peak-Load Pricing and Optimal Capacity: Comment," *American Economic Review,* Vol. 58 (1968), pp 168–70.
5. ———. "The Optimality of Pure Competition in the Capacity Problem: Further Comment," *Quarterly Journal of Economics,* Vol. 83 (1969), pp. 341–43.
6. De Salvia, D. N. "An Application of Peak-Load Pricing," *Journal of Business,* Vol. 42 (1969), pp. 458–76.
7. Dreze, J. A. "Some Postwar Contributions of French Economists to Theory and Public Policy," *American Economic Review* (*Supplement*), Vol. 52 (1964), pp. 1–64.
8. Henderson, J. and Quandt, R. *Microeconomic Theory.* New York: McGraw-Hill, 1958.
9. Hicks, J. R. *Value and Capital.* Oxford: Oxford University Press, 1939.
10. ———. "The Rehabilitation of Consumer's Surplus," *Review of Economic Studies,* Vol. 8 (1940–41), pp. 108–116.
11. ———. "The Four Consumer's Surpluses," *Review of Economic Studies,* Vol. 11 (1943), pp. 31–41.
12. Hirshleifer, J. "Peak-Loads and Efficient Pricing: Comment," *Quarterly Journal of Economics,* Vol. 72 (1958), pp. 451–68.
13. Hotelling, H. "Demand Functions with Limited Budgets, "*Econometrica,* Vol. 3 (1935), pp. 66–78.
14. ———. "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions," *Journal of Political Economy,* Vol. 40 (1932), pp. 577–616.
15. Marshall, A. *Principles of Economics.* 8th rev. ed. London: MacMillan and Company, 1925.
16. Nelson, J. R., ed. *Marginal Cost Pricing in Practice.* Englewood Cliffs, N. J.: Prentice Hall, 1964.
17. Officer, L. H. "The Optimality of Pure Competition in the Capacity Problem," *Quarterly Journal of Economics,* Vol. 80 (1966), pp. 647–51.
18. Pfouts, R. W. "A Critique of Some Recent Contributions to the Theory of Consumer's Surplus," *Southern Economic Journal,* Vol. 19, No. 3 (January 1953), pp. 315–33.

---

[23] See the first term on the right-hand side of equation (1).

19. Phillips, C. F. *The Economics of Regulation.* Homewood, Illinois: Richard D. Irwin, Inc., 1965.
20. Samuelson, P A. *Foundations of Economic Analysis.* New York: Atheneum, 1965.
21. Schultz, H. "Interrelations of Demand," *Journal of Political Economy,* Vol. 41 (1933), pp. 468–512.
22. Steiner, P. "Peak-Loads and Efficient Pricing," *Quarterly Journal of Economics,* Vol. 71 (November 1957), pp. 585–610.
23. Turvey, R. "Peak-Load Pricing," *Journal of Political Economy,* Vol. 76 (1968), pp. 101–112.
24. ———. "Marginal Cost," *The Economic Journal,* Vol. 79 (1969), pp. 282–99.
25. Wellisz, S. "Regulation of Natural Gas Pipeline Companies: an Economic Analysis," *Journal of Political Economy,* Vol. 71 (1963), pp. 30–43.
26. Williamson, O. E. "Peak-Load Pricing and Optimal Capacity Under Indivisibility Constraints," *American Economic Review,* Vol. 56 (1966), pp. 810–27.
27. Winch, D. "Consumer's Surplus and the Compensation Principle," *American Economic Review,* Vol. 55 (1965), pp. 395–423.

Regulation of motor freight transportation: a quantitative evaluation of policy

James Sloss

Research Consultant
The Transportation Center at Northwestern University

This study develops a way of using data to measure certain economic effects of transport regulation, particularly the rate effects associated with regulation of the motor carrier industry. The comprehensiveness of regulation in the United States has made cross-section studies of the effects of regulation virtually impossible. However, province-to-province differences in regulation in Canada produce statistical information on the for-hire motor truck industry making it possible to estimate the rate levels associated with particular degrees of regulation there. A series of regression equations indicate that trucking charges in Canada have been higher under regulation than in its absence. Extension of the analysis to the for-hire motor trucking industry in the United States, using the more limited data available here, suggests that the same conditions may account for increased motor freight charges of $300 million per annum.

■ The past decade has witnessed a continuing debate between supporters of the existing scheme of economic regulation of transportation, on the one hand, and those favoring substantial modifications (generally in the direction of lessened government controls), on the other. The arguments advanced by both sides have been based largely on inferences and allegations; little empirical evidence has been introduced either by those proposing continuance of the *status quo* or by those who would substitute greater reliance on competition for much of the prevailing regulatory structure. This study seeks to alleviate this deficiency to some degree by using quantitative information to assess the economic effects from government regulation of the suppliers of trucking transport services.

1. Regulation of motor transport in the United States

■ Events leading to federal regulation of the trucking industry in the U. S.,[1] which culminated in the *Motor Carrier Act* of 1935, began with

---

Mr Sloss holds the B A. degree, *summa cum laude*, from Harvard College and the MBA degree from the University of Chicago. He has been with Northwestern University's Transportation Center since 1963. Since 1959 he has been a Lecturer in Transportation Administration at Northwestern. He also served as Assistant Director of the University of Chicago's "Downtown Program" from 1961 to 1963 Mr. Sloss' career in teaching and research began after over 20 years in business

The author gratefully acknowledges the assistance of Ashish Sen, Assistant Professor of Urban Sciences, University of Illinois, Chicago Circle campus, in providing invaluable advice and guidance in the selection and development of statistical and mathematical procedures. Many other individuals provided assistance without which this study would have been impossible Among them are Professors Paul W. MacAvoy of the Massachusetts Institute of Technology, George W Wilson of Indiana University, Richard B. Heflebower of Northwestern University, and Lars Kritz of the Industrial Institute for Economic and Social Research in Stockholm, William T. Olsen, Florida State University; Dr J. A Bailey, Director of the Transportation Center at Northwestern University, Benjamin H. Jacobson, Librarian of the Transportation Center Library, and members of his staff; J. P. Worshill of the Canadian Pacific (Chicago), officials of the Dominion Bureau of Statistics, the Provincial Transport Ministries, Regulating Boards and Commissions, and motor carrier tariff bureaus.

[1] See, for example, [32], [35], and [37].

a bill introduced in Congress in 1909 "for the regulation, identification and registration of motor vehicles engaged in interstate travel."[2] But the major impetus for federal regulation of motor carriage followed a 1925 decision of the U. S. Supreme Court, which held that a state lacked authority to regulate common-carrier motor vehicles engaged in interstate transportation "for controlling competition on ground of public convenience."[3] Between that date and enactment of the *Motor Carrier Act* ten years later two significant events precipitated affirmative action by Congress.

The first was a report by the Federal Coordinator of Transportation, Joseph B. Eastman. Under the terms of the statute creating the office of coordinator, Mr. Eastman was required to study and report to Congress on the need for "federal legislation to regulate other transportation agencies and to promote the proper coordination of all means of transport." His recommendation that interstate motor carrier operations be placed under the control of the Interstate Commerce Commission was one of the crucial factors leading to passage of the 1935 *Motor Carrier Act.*

The second event was the Supreme Court's decision of May 27, 1935, in the "Schechter" case,[4] declaring the *National Recovery Act* unconstitutional. This terminated a code under which the trucking industry was attempting to limit the forces of competition and stabilize rates. The *Motor Carrier Act* became law on August 9, 1935, and there were no important changes in public policy for the next twenty years.

Arguments on government regulation took a new turn in 1955 with the publication of the "Weeks Report," which advocated amending federal statutes "to permit greater reliance on competitive forces in transportation pricing."[5] The specific recommendations for implementing this appeared to be designed to aid the railroads *vis-à-vis* other modes of transport, however, so that strong opposition to the proposed changes developed among the motor and water carriers. Their opposition continues.

Reaffirmation of the basic tenets enunciated in the Weeks Report by Presidents Kennedy, Johnson, and Nixon has produced few results other than volumes of Congressional hearings on modifications of federal transport regulation. Views favoring less restrictive controls over entry and pricing in transportation have been expressed by the Council of Economic Advisors and by past and present members of the Department of Transportation.[6] The American Trucking Associations and members of the Interstate Commerce Commission have firmly opposed any relaxation of motor carrier regulation.[7]

While federal regulation of interstate transport developed under the Constitution's "Interstate Commerce Clause,"[8] regulation by

---

[2] H. R. 5176. See Wagner [35], p 93.

[3] *Buck v Kuykendall*, 266 U. S. 307 (1925).

[4] *Schechter v. United States*, 295 U. S. 495 (1935).

[5] See [36].

[6] See, for example, Mackey [18]. Also, see the remarks of J. M. Beggs, Under Secretary of Transportation, as reported in *Traffic World*, July 5, 1969, p. 33, and a speech by Paul W Cherington, former Assistant Secretary of Transportation for Policy and International Affairs, delivered before the Traffic Club of New York City, reported in *Traffic World*, November 22, 1969, pp. 24 and 25.

[7] See, for example, [2], and [26], also, remarks of Commissioner Dale W. Hardin as reported in *Traffic World*, March 14, 1970, p. 33.

[8] *U. S. Constitution*, Art. I. Sec. 8.

TABLE 1

TRUCK TRAFFIC ON MAIN RURAL ROADS, UNITED STATES, 1955

PERCENTAGE DISTRIBUTION OF TON–MILES BY SINGLE–UNIT TRUCKS AND COMBINATIONS, INTERSTATE AND INTRASTATE, BY CENSUS REGIONS. FOR–HIRE CARRIERS WITH I C.C. AUTHORIZATIONS AND OTHER FOR–HIRE CARRIERS										
	UNITED STATES	NEW ENGLAND	MIDDLE ATLANTIC	SOUTH ATLANTIC	EAST NORTH CENTRAL	EAST SOUTH CENTRAL	WEST NORTH CENTRAL	WEST SOUTH CENTRAL	MOUNTAIN	PACIFIC
PERCENT INTERSTATE TON–MILES	63.21	55.35	68.42	73.40	75.98	69.78	56.16	55.66	68.61	29.83
PERCENT INTRASTATE TON–MILES	36.79	44.65	31.58	26.60	24.02	30.22	43.84	44.34	31.39	70.17

SOURCE: INTERSTATE COMMERCE COMMISSION, BUREAU OF TRANSPORT ECONOMICS AND STATISTICS, "TRUCK TRAVEL ON MAIN RURAL ROADS, 1955", STATEMENT NO. 5710, FILE NO. 51–D–4, WASHINGTON, JULY 1957.

the individual states of commercial motor traffic within their own borders was evolving along similar lines. The percentage of total intercity truck traffic that does not cross a state line or international border, measured in ton-miles, has been relatively small (as Table 1 indicates).[9] Nevertheless, every state except Delaware and New Jersey has adopted economic controls over intrastate trucking comparable to those governing interstate services in Part II of the *Interstate Commerce Act*. "Many of the state laws and amendments adopted since 1935 follow the general outline of the Federal Act . . . . [A]lso, the patterning of regulations after the comprehensive *Motor Carrier Act* has made state regulation more effective."[10]

During the development of regulation on the state and federal levels it has been difficult to assess the associated economic consequences. Estimates of the effects on prices for trucking services have been published for various agricultural commodities which, following I.C.C. or court rulings, have been added to or removed from the list of "property" exempt from federal regulation.

One such study, involving fresh-dressed and frozen poultry which had been added to the list of exempted agricultural products by court order, found that truck rates into 12 major markets were reduced 33 percent and 36 percent, respectively, after federal regulation was removed.[11] Subsequent studies by the U. S. Department of Agriculture confirmed declines in truck rates on fresh processed poultry ranging from 12 to 53 percent. For frozen poultry, reductions ranged from 16 to 59 percent when these items were exempted from federal regulation. According to the Department, interstate

---

[9] It was necessary to rely on 1955 data in support of this assertion since more recent distributions of interstate and intrastate truck movements could not be found. See also [17]

[10] In [17], p 42 See also [10].

[11] Cf [19], and [29].

TABLE 2

CANADIAN FOR–HIRE MOTOR TRANSPORT TRAFFIC

PERCENTAGE DISTRIBUTION OF TON–MILES, EXTRAPROVINCIAL/INTERNATIONAL AND INTRAPROVINCIAL, BY CONTINENTAL PROVINCES, 1958 AND 1963

		CANADA*	NOVA SCOTIA	NEW BRUNSWICK	QUEBEC	ONTARIO	MANITOBA	SASKATCHEWAN	ALBERTA	BRITISH COLUMBIA
PERCENT EXTRAPROVINCIAL/ INTERNATIONAL TON–MILES	1958	39.8	4.5	36.8	42.6	33.0	73.9	24.6	50.3	35.8
	1963	43.3	14.5	42.5	31.1	25.2	71.4	38.2	66.6	51.6
PERCENT INTRAPROVINCIAL TON–MILES	1958	60.2	95.5	63.2	57.4	67.0	26.1	75.4	49.7	64.2
	1963	56.7	85.5	57.5	68.9	74.8	28.6	61.8	33.4	48.4

SOURCE: DOMINION BUREAU OF STATISTICS, PUBLIC FINANCE AND TRANSPORTATION DIVISION, TRANSPORTATION AND PUBLIC UTILITIES SECTION, *MOTOR TRANSPORT TRAFFIC*, CATALOGUE NOS. 53-207 THROUGH 53-214, INCLUSIVE, OTTAWA, QUEEN'S PRINTER AND CONTROLLER OF STATIONERY, ANNUAL

* EXCLUSIVE OF THE PROVINCES OF NEWFOUNDLAND AND PRINCE EDWARD ISLAND, THE NORTHWEST AND YUKON TERRITORIES.

trucking of these commodities under the exemption has been accompanied by improved service as well as lower rates.[12]

Overall, the question is whether lower charges at an acceptable level of service have characterized that segment of the industry exempt from economic regulation. There is not enough information to permit reliable, meaningful answers. Fortunately, the trucking industry of Canada serves as more fertile ground for a comprehensive investigation.

2. Truck regulation in Canada

■ The geographical and political structures of Canada have influenced the regulation of the Canadian trucking industry significantly. The average area of the eight mainland provinces (omitting Newfoundland and Prince Edward Island) is 272,683 square miles, about quadruple the 64,225-square-mile average of the 48 continental states of the U. S. Furthermore, the population of Canada is highly concentrated. According to the 1961 census, 65 percent of the mainland population inhabited the provinces of Quebec and Ontario. Thus, it is not surprising that intraprovincial motor freight traffic has exceeded the traffic crossing provincial or international boundaries, as Table 2 indicates.[13]

Political conditions also differ greatly from those in the United States. In contrast to the relatively uniform federal and state regulatory systems applicable to the U. S. trucking industry, the pattern of economic controls in Canada has been marked by diversity, both among the individual provinces and in the relationship of the Dominion government to the provincial regulatory structure. This diversity makes it possible for us to compare the charges imposed by trucking firms subject to regulation with those of firms operating in a relatively uncontrolled environment.

---

[12] In [29], pp. 1–3.

[13] These respective distributions are based on ton-miles for 1958 and 1963, the years spanning the period utilized in the quantitative analysis presented in later sections of this article.

In the *British North American Act* of 1867, which created the Dominion of Canada, authority to control navigation and shipping, railways, canals, and telegraphs was vested in the national Parliament. The Act contained no specific reference to highway traffic, so provinces evolved controls for this newer form of transport. The development in this century of an extensive network of hard-surfaced roads, combined with technological advances in vehicle operation, brought long-distance trucking into competition with the railways and raised the issue as to which jurisdiction had authority to regulate such movements. The question was settled by a 1952 decision of the Canadian Supreme Court,[14] sustained on appeal by the British Privy Council, pronouncing the right of the Federal Parliament to regulate extraprovincial and international highway transport services.[15]

The Parliament in Ottawa responded to these decisions by enacting the *Motor Vehicle Transport Act* of 1954, empowering the transport boards of the individual provinces to license trucking firms for extraprovincial operation and to determine and regulate tariffs and tolls charged for extraprovincial truck transport "as if . . . [it] were local transport." By its terms, the Act took effect in a province by an appropriate proclamation of the Governor in Council. By 1961 the Act had been proclaimed in every province except Newfoundland.

Since the 1954 statute authorized each province to apply the same controls to extraprovincial trucking as were imposed on "local transport," widely diverse regulatory practices were applied by the provinces. The differences were so great, in fact, that no common scope and effect of extraprovincial truck regulation were achieved during the period extending from 1958 through 1963—the years for which data was used in this study.

Numerous practical difficulties arose as a consequence of this legislation. A provincial board could, for example, control entry of extraprovincial carriers by the simple expedient of refusing to issue the requisite license or certificate. Since the 1954 statute made no provision for joint hearings by provincial boards, an applicant might have to make separate appearances before as many as five boards.[16] Under these circumstances, control of rates on extraprovincial movements was unwieldy as well. According to a Royal Commission reporting in 1955, it was agreed that provincial boards should not attempt to control rates charged by carriers engaging in extraprovincial operations.[17] The final step was taken with passage of Part III of the *National Transportation Act* of 1967. When this goes into effect by proclamation as of May 15, 1970, regulation of extraprovincial trucking will be transferred to the Canadian Transport Commission.[18]

---

[14] *Israel Winner, doing business under the name and style of MacKenzie Coach Lines, (Defendant) Appellant, against S. M. T. (Eastern) Ltd., (Plaintiff) Respondent.* Canada Law Reports, Supreme Court of Canada (SCR) 1951, p. 887.

[15] *Attorney General for Ontario et al v. Winner et al.; Winner et al. v. S. M T. (Eastern) Limited et al*, Her Majesty's Privy Council, February 22, 1954, All England Law Reports, October 21, 1954.

[16] See Dartnell [13], p 137, and Carr et al [9], p. 39, fn 2.

[17] See Turgeon [28], p. 24

In contrast to the overlapping jurisdiction over extraprovincial trucking, there has been no duplication of authority to regulate service within the borders of a single province. The regulation of such traffic has been the prerogative of the provincial and territorial governments. Some control over entry into the intraprovincial trucking business began to be exercised as early as 1927 in one province, and as recently as 1968 in another. Mandatory observance of rates filed with a provincial board commenced in 1930 in one instance, while in at least two provinces no government intervention in the pricing of trucking services had yet been authorized by 1969. This diversity has provided a favorable set of conditions for weighing the effects of regulation on the level of trucking charges, simply by comparing the performance of trucking companies, province-by-province, according to the prevailing degrees of regulation.

A word of caution may be appropriate concerning the use of this seemingly near-ideal situation. While the variations in authorities and enforcement procedures offer an opportunity for evaluating the costs of regulation, they also create complications. The structure of this study required categorizing each province as either "regulating" or "non-regulating." No such simple distinction exists for many of the provinces, of course. Furthermore, classification was complicated by inconsistencies among the regulators' own descriptions of their administrative procedures. This study, therefore, is a first endeavor to take advantage of diverse regulatory practices among the provinces, since it is necessarily affected by imprecise definition of activities.

The principal source of information used was the Canadian Dominion Bureau of Statistics. From 1957 to 1964 the Bureau issued a series of statistical reports on the performance of for-hire trucking companies.[19] Because of the usual reporting and sampling errors associated with a new system of data collection, it was considered inadvisable to use the first of these. Also, figures from the final report in the series were omitted on the strength of information in its introduction explaining changes in reporting methods and declaring: "The 1964 figures shown under various traffic headings are not entirely comparable to the statistics for 1963."[20] Accordingly, the 1958-1963 data were analyzed for the effects of regulation.

The Dominion Bureau of Statistics reported elsewhere such items as annual license fees, taxes, and motor carrier regulations by provinces. Excerpts covering provincial trucking regulations for the years 1959 and 1963 have been reproduced in the appendix in Tables 14 and 15.[21] These indicate that entry was controlled in all provinces except Newfoundland and that changes in rate regulation during the period of the study occurred only in New Brunswick and

---

[18] See *Traffic World*, June 6, 1970, pp. 22–23.

[19] See [5], Catalogue No 53–207 designated National Estimates, and Catalogue Nos 53–208 through 53–214 designated Provincial Estimates See years 1957–1964.

[20] See [5], 1964, Introduction, pp. 8 and 10. Reasons for discontinuing this report series have been supplied in the preface to Steeves [27]

[21] See [6], special supplement issued October 1, 1960, catalogue No 53–203. See [6] for 1963, Part 1, table 4, pp. 18–19, catalogue No. 53–217.

TABLE 3

STATUS OF PROVINCIAL TRUCK REGULATION REPORTED BY SIX SOURCES PUBLISHED BETWEEN 1956 AND 1967

NAME OF SOURCE AUTHOR PUBLISHER DATE PUBLISHED	PRINCE EDWARD ISLAND		NOVA SCOTIA		NEW BRUNSWICK		QUEBEC		ONTARIO		MANITOBA		SASKAT-CHEWAN		ALBERTA		BRITISH COLUMBIA	
	E	R	E	R	E	R	E	R	E	R	E	R	E	R	E	R	E	R
TRANSPORTATION IN CANADA, APPENDIX A J–C. LESSARD QUEEN'S PRINTER 1956	NO	NO	NO	NO	YES	NO	YES	YES	YES	NO	YES	YES	YES	YES	NO	NO	YES	YES
THE MOTOR VEHICLE D B S (1959 DATA) QUEEN'S PRINTER 1960	YES	YES	YES	NO	YES	NO	YES	YES	YES	NO	YES	NO	YES	YES	YES	NO	YES	YES
CANADIAN TRAFFIC & TRANSP'N. MANAGEMENT, VOL. II, SUBJECT 25. CAN IND. TRAF. LGE 1960	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	NO	NO	YES	YES
TRUCK–RAIL COMPETITION IN CANADA. D.W. CARR & ASSOC. QUEEN'S PRINTER 1962	–	LTD.	–	LTD	–	LTD	–	YES	–	NO	–	LTD	–	LTD.	–	NO	–	LTD
THE MOTOR VEHICLE PART I, TABLE 4 D.B.S. (1963 DATA) QUEEN'S PRINTER 1964	YES	YES	YES	NO	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES	NO	YES	YES
CANADIAN TRANSP'N ECONOMICS, CH 18 A W. CURRIE UN. OF TORONTO PRESS 1967	YES	NO	YES	NO	YES	FILE ONLY	YES	YES	YES	FILE ONLY	YES	YES	YES	YES	NO	NO	YES	YES

E = CONTROL OF ENTRY IS REGULATED R = RATES ARE REGULATED. LTD = LIMITED RATE REGULATION.
ALL SOURCES AGREE NEITHER ENTRY NOR RATES ARE REGULATED IN NEWFOUNDLAND.

Manitoba. New Brunswick, listed as not exercising its power over filing and regulating freight rates in 1959, was reported to be doing so in 1963; Manitoba, shown in 1959 as lacking power to regulate rates, was shown in the 1963 report to have such power and to be exercising it.[22]

The findings of the Dominion Bureau of Statistics have been matched with the conclusions of four other authoritative sources; these appear in Table 3. Some inconsistencies can be discerned, mostly in relation to the Maritime Provinces of Prince Edward Island, Nova Scotia, and New Brunswick.[23] With respect to entry control, all sources except the Dominion Bureau of Statistics (and D. W. Carr, who provided no entry information) agreed that entry into intraprovincial trucking in Alberta was not regulated.[24] With regard to rates and charges, there is a general consensus on where and when regulation existed. The only exceptions are in the Maritime Provinces and Manitoba. The preponderant evidence now shows the Dominion Bureau erred in reporting in 1959 that the Motor Carrier Board of Manitoba lacked power to regulate intraprovincial rates. The other inconsistencies can be attributed to changes in provincial regulatory policies. This might have been the case for New Brunswick, where an important amendment took effect in 1961, and for Ontario, which inaugurated filing of motor carrier rates with its Highway Transport Board in 1963.

---

[22] According to all other sources consulted, Manitoba had granted power to its Motor Carrier Board to regulate intraprovincial truck rates as early as 1956. In 1959 the Board was actually prescribing the rates to be charged on intraprovincial truck traffic.

[23] Written communications from persons closely identified with the regulatory practices in these provinces have served to clarify these apparent disagreements, as may be noted by referring to Appendix I, page 356

In order to obtain information directly on the administration of the provincial statutes, a questionnaire was prepared and distributed to each provincial board or commission. The questionnaire asked for information on the nature and effective dates of entry and rate controls imposed on common and contract trucking firms operating intraprovincial and extraprovincial services. Responses were received from the regulating agencies of all ten provinces as well as from two trucking industry associations.[25] Responses covering extraprovincial operations are summarized in Table 4, and those covering intraprovincial movements of freight are in Table 5. So few differences between the regulation of common and contract carriers were noted on the forms that they have been combined in both tables. Amplifying comments received with some of the questionnaire returns have been reproduced in the appendix.

Separate details for intraprovincial and extraprovincial regulations were requested on the questionnaires; responses from the six eastern provinces did not differentiate between these categories. The four western provinces did report certain dissimilarities, which can be seen by comparing Tables 4 and 5. In general, descriptions of intraprovincial regulation agreed substantially with status reports derived from the other sources listed in Table 3. Clarification of the status of extraprovincial trucking regulation was not measurably enhanced by the questionnaire responses. It would appear, therefore, that the classification of intraprovincial trucking services as "regulated" or "not regulated" should have been relatively easy.

In point of fact, classification was anything but easy. Variations in the phraseology of the provincial statutes, in combination with the diverse interpretations and degrees of enforcement of these statutes by the regulating authorities, have made the allocation process quite judgmental.

The first step toward designating the provinces as "regulating" or "non-regulating" has been to formulate a definition of the term "regulation" in this context. Definitions of the term applied to the transportation industries are as diverse as the authors who have written on the subject. There has been, however, general agreement that government regulation of privately owned transportation firms evolved because imperfections in the structure of markets for transport services resulted in either too little or too much competition. In the case of the trucking industry, the latter has provided the rationale for placing it under government controls.[26] Regulation of motor transport in Canada has been justified on the grounds that it would prevent the "wastes" of unrestrained competition and ensure adequate service, fair wages, responsible management and greater stability in rates.[27]

---

[24] This conclusion has been confirmed by an examination of a lengthy series of annual reports of the Department of Highways of the Province of Alberta, Motor Vehicles Branch, which include reports of the provincial regulating agency, the Highway Traffic Board. See [1], pp. 56–61. Agreement as to the nonexistence of entry controls in Newfoundland was unanimous.

[25] The Alberta Motor Transport Association, Canadian Transport Tariff Bureau Association, and the Niagara Frontier Tariff Bureau, Inc., also furnished enlightening information concerning the regulatory status of various categories of Canadian and international motor freight traffic. See Appendix 1.

It has been argued that effective regulation requires that both the supply of a service and its price be placed under the supervision of a public agency. Citing one such declaration: "In all regulated industries . . . an entrant must apply to the relevant commission for a 'certificate of public convenience and necessity.' By denying certificates, it is thought, a commission can prevent speculative construction, over-capacity, and ruinous competition, thus assuring financial stability and continued service *at regulated rates*."[28] Another author has described the critical areas of regulation as encompassing control of prices and the restriction of rights to enter and withdraw from such industries as transportation, public utilities, and communications.[29] Failure to regulate either entry or price, due to legislative policy or administrative discretion, justifies here the designation of "non-regulation" for the industry and the province.

The second step is to make the best possible reconciliation of the views concerning the status of truck regulation in each Canadian province in the 1958–1963 period. It has been generally agreed that four provinces maintained effective regulatory controls over intra-provincial motor freight service during the relevant period. There is ambiguity as to the status of some of the rest of the provinces; on balance, the remaining six were considered to lack the requisite legislative authority, or else to have refrained from enforcement of powers granted by their respective parliaments. The assignment of the Canadian provinces is, then, as follows:

"Regulating" Provinces	*"Non-Regulating" Provinces*
Quebec	Newfoundland
Manitoba	Prince Edward Island
Saskatchewan	Nova Scotia
British Columbia	New Brunswick
	Ontario
	Alberta

New Brunswick and Ontario amended their regulatory statutes between the beginning and end of the period covered here. In September 1961 the laws of New Brunswick were altered to provide that for-hire truckers "shall (not) accept goods for carriage or carry goods at any rate other than that approved by the Board." In May 1963, Ontario introduced a requirement that all for-hire trucking rates and charges be filed with the provincial Highway Transport Board. In neither case were these modifications deemed sufficiently

---

[26] See [21], pp. 486–487.

[27] See [12], p. 454.

[28] See [38], p. 285 ff, also, [39], pp. 10–32.

[29] See [20], pp. 257–8, and [21], p. 4. "As commonly used. . . the phase *the regulated industries* refers to a diverse group of businesses which have been subjected to detailed local, state, and federal regulation as to rates and services."

TABLE 4

CANADIAN FOR–HIRE TRUCKING QUESTIONNAIRE RESPONSES: EXTRAPROVINCIAL CARRIERS

	NEWFOUNDLAND	PRINCE EDWARD ISL.	NOVA SCOTIA	NEW BRUNSWICK	QUEBEC
CERTIFICATE OF PUBLIC CONVENIENCE & NECESSITY REQUIRED?	(MOTOR VEHICLE TRANSPORT ACT) OPERATING LICENSE	NO	NO	YES	CERTIFICATE OF PUBLIC CONVENIENCE & NECESSITY IS PART OF THE PERMIT TO ENGAGE IN TRUCKING SERVICE
PERMIT TO ENGAGE IN SERVICE REQUIRED?	YES {VEHICLE PERMIT}	YES	LICENSE – YES	YES	
IF CERTIFICATE AND/OR PERMIT REQUIRED, WHEN DID THIS TAKE EFFECT?	JUNE 1st, 1968	1961	LICENSE REQUIRED SINCE 1959	YES	1930
CLASSIF., RATES, RULES, REGULATIONS REQUIRED TO BE PUBLISHED, FILED WITH REGULATING AGENCY & KEPT OPEN FOR PUBLIC INSPECTION?	REQUIRED TO BE FILED FOR BOARD'S APPROVAL	YES [(a)]	FILING ONLY	YES	YES
IF PUBLICATION REQUIRED, WHEN DID THIS TAKE EFFECT?	N/A	N/A	N/A	N/A	1930
ARE PUBLISHED RATES, RULES, & REGULATIONS REQUIRED BY LAW TO BE OBSERVED?	N/A	YES [(a)]	YES	YES	YES
ARE PUBLISHED RATES OBSERVED IN FACT?	N/A	YES [(a)]	GENERALLY – YES (b)	YES	95%
IF OBSERVANCE OF PUBLISHED RATES REQUIRED, WHEN DID THIS TAKE EFFECT?	N/A	1961 [(a)]	1959 [(b)]	N/A	1948
ARE MOTOR CARRIER TARIFFS PUBLISHED BY RATE BUREAUS FOR MEMBER FIRMS?	NO	YES	FOR FURNITURE MOVERS ONLY	YES (COMMON CAR.) NO (CONTRACT CAR.)	SOME ARE
DO MEMBER CARRIERS GENERALLY ADHERE TO RATES, CLASSIF. & RULES PUBLISHED BY BUREAU?	N/A	YES	YES, WHERE APPLICABLE (SEE ABOVE)	YES (COMMON CAR.) N/A CONTRACT CAR.)	YES

(a) THE WORD "PUBLISHED" WAS STRICKEN

(b) "FILED" SUBSTITUTED FOR "PUBLISHED"

(c) "THERE ARE, OF COURSE, ALWAYS EXCEPTIONS"

(d) "STATUTORY" SUBSTITUTED FOR "PUBLISHED"

ONTARIO	MANITOBA	SASKATCHEWAN	ALBERTA	BRITISH COLUMBIA
YES	YES	YES	PROPER REGISTRATION OF VEHICLE PLUS CERTIFICATE REQUIRED.	NO
YES	YES	YES	–	YES
1927	1954	HAS BEEN IN EFFECT FOR MANY YEARS	WE COMMENCED IN 1957 AND FULLY DEVELOPED IN 1958	1939
YES	NO	NON-APPLICABLE	NO RATE FILING REQUIRED	NO
1963	–	NON-APPLICABLE	N/A	-
YES	(d) NO	NON-APPLICABLE	RULES & REGULATIONS REQUIRED TO BE OBSERVED UNDER PUBL. SERV. VEHICLES ACT	NO
(c) YES	NOT REQUIRED (d)	NON-APPLICABLE	IN SOME CASES, BY MUTUAL AGREEMENT AMONG CARRIERS	NOT REGULATED
1963	–	NON-APPLICABLE	NO LEGAL PROVISION FOR OBSERVANCE OF PUBLISHED RATES	—
YES	YES	NON-APPLICABLE	WESTERN TARIFF BUREAU; CANADIAN TRANSPORT TARIFF BUREAU (e)	N/A
(c) YES	NOT REQUIRED	NON-APPLICABLE	THIS IS NOT KNOWN (f) YES & NO (e)	NOT REGULATED

(e) CITING ALBERTA MTR. TRANSP. ASS'N.

(f) CITING ALBERTA HIGHWAY TRAFFIC BD.

TABLE 5

CANADIAN FOR–HIRE TRUCKING QUESTIONNAIRE RESPONSES. INTRAPROVINCIAL CARRIERS

	NEWFOUNDLAND	PRINCE EDWARD IS.	NOVA SCOTIA	NEW BRUNSWICK	QUEBEC
CERTIFICATE OF PUBLIC CONVENIENCE & NECESSITY REQUIRED?	YES	NO	NO	YES	CERTIFICATE OF PUBLIC CONVENIENCE & NECESSITY IS PART OF THE PERMIT TO ENGAGE IN TRUCKING SERVICE
PERMIT TO ENGAGE IN SERVICE REQUIRED?	YES {VEHICLE PERMIT}	YES	LICENSE – YES	YES	
IF CERTIFICATE AND/OR PERMIT REQUIRED, WHEN DID THIS TAKE EFFECT?	JUNE 1st, 1968	1961	LICENSE REQ'D SINCE 1959	R.S.N.B. 1937 AS AMENDED IN 1952 AND 1957	1930
CLASSIF., RATES, RULES, REGULATIONS REQUIRED TO BE PUBLISHED, FILED WITH REGULATING AGENCY & KEPT OPEN FOR PUBLIC INSPECTION?	REQUIRED TO BE FILED FOR BOARD'S APPROVAL	YES (g)	FILING ONLY	RATES REQUIRED TO BE FILED REGULATION 49	YES
IF PUBLICATION REQUIRED, WHEN DID THIS TAKE EFFECT?	N/A	N/A	N/A	N/A	1930
ARE PUBLISHED RATES, RULES, & REGULATIONS REQUIRED BY LAW TO BE OBSERVED?	N/A	YES (g)	YES	YES	YES
ARE PUBLISHED RATES, OBSERVED IN FACT?	N/A	YES (g)	GENERALLY (h) YES	YES	95%
IF OBSERVANCE OF PUBLISHED RATES REQUIRED, WHEN DID THIS TAKE EFFECT?	N/A	1961 (g)	1959 (h)	N/A	1948
ARE MOTOR CARRIER TARIFFS PUBLISHED BY RATE BUREAUS FOR MEMBER FIRMS?	NO	YES	FOR FURNITURE MOVERS ONLY	YES (COMMON CAR.) NO (CONTRACT CAR.)	SOME ARE
DO MEMBER CARRIERS GENERALLY ADHERE TO RATES, CLASSIF. & RULES PUBLISHED BY BUREAU?	N/A	YES	YES, WHERE APPLICABLE (SEE ABOVE)	YES (COMMON CAR) N/A (CONTRACT CAR.)	YES

(g) THE WORD "PUBLISHED" WAS STRICKEN

(h) "FILED" SUBSTITUTED FOR "PUBLISHED"

(i) "THERE ARE, OF COURSE, ALWAYS EXCEPTIONS"

(j) "STATUTORY" SUBSTITUTED FOR "PUBLISHED"

ONTARIO	MANITOBA	SASKATCHEWAN	ALBERTA	BRITISH COLUMBIA
YES	YES	YES	NO PUBLIC CONVENIENCE & NECESSITY NEED BE PROVED	NO
YES	YES	YES	PROPER VEHICLE REGISTRATION IS REQUIRED.	YES
1927	1934	HAS BEEN IN EFFECT FOR MANY YEARS.	–	1939
YES	NO	YES	NO RATE FILING REQUIRED	YES
1963	–	IN EFFECT FOR MANY YEARS.	N/A	1939
YES	YES (j)	YES	RULES & REGULATIONS REQUIRED TO BE OBSERVED UNDER PUBLIC SERVICE VEHICLE ACT	YES
YES (i)	PARTLY (j)	YES	IN SOME CASES, BY MUTUAL AGREEMENT AMONG CARRIERS.	SUBSTANTIALLY YES
1963	–	WITHIN RECENT YEARS.	NO LEGAL PROVISION FOR OBSERVANCE OF PUBLISHED RATES.	1939
YES	NO	PUBLISHED BY SASK. TRK'G. ASS'N FOR ITS MEMBERS.	ALBERTA SHIPPERS GUIDE, WESTERN TARIFF BUREAU (k)	SOME CARRIERS FILE THROUGH A TARIFF AGENT, OTHERS FILE ON THEIR OWN. (m)
YES (i)	A GOOD QUESTION PARTLY	YES	THIS IS NOT KNOWN (l) YES & NO (k)	SUBSTANTIALLY YES

(k) CITING ALBERTA MTR. TRANSP. ASS'N.

(l) CITING ALBERTA HIGHWAY TRAFFIC BD.

(m) MOST CONTRACT CARRIERS FILE THEIR OWN CONTRACTS consequential to change the assigned regulatory status, owing largely to amplifying statements (which are set forth in the Appendix).

This assignment was found in retrospect to have created anomalies in the case of Quebec. Initial quantitative analyses produced results for this province which differed markedly from *a priori* expectations and from results calculated for the other provinces which had been classified "Regulating."[30] At the same time, there was no question about the Quebec Transportation Board's possession of the requisite statutory authority to regulate motor freight transport; broad powers had been delegated to the Board in the *Transportation Board Act* over the rates to be charged for transport service within the province and over the granting of certificates to engage in a transport undertaking or to expand an existing operation. One explanation of Quebec's behavior was the disclosure that in 1960 the motor carriers of that province had inaugurated "agreed charges" pricing, a practice introduced by the Canadian railways under the aegis of the *Transport Act* of 1938.[31] Each agreed charge was negotiated directly by an individual shipper with a particular motor carrier. In return for a rate below that usually applicable to the commodity to be carried, the shipper agreed to direct a fixed volume of traffic through the carrier party to the agreement. Only in Quebec have "agreed charges" been a significant factor in the total highway rate structure; it was estimated that about 30 percent of all highway freight moved in the province under some 200 agreed charge tariffs prior to the Quebec Transporation Board's decision, effective June 16, 1969, requiring renegotiation and refiling of all outstanding agreed charges between shippers and motor carriers. Despite the absence of precise figures to indicate the extent of reductions attributable to "agreed charges," they evidently resulted in average charges in Quebec below those for comparable service in other regulated provinces. Efforts by the organized trucking industry to secure the elimination of agreed charges tend to support this view.[32]

Regulation of agreed charges by the Quebec Transportation Board appears to have been confined to an insistence that all shippers have access to privileges provided in a given contract and that all carriers authorized to give the same service should participate on demand of the shipper.[33]

The Annual Reports of the Quebec Transportation Board attest to the Board's adherence to a policy of unusual liberality in its regulation of trucking, both in rates and in permitting entry of new firms and expansion of service by existing firms. These reports disclose a relatively light case load for the Board's investigators in reviewing the rates of trucking firms and a high percentage of approvals in proceedings involving permit applications, as noted in the Appendix in Table 16. Professor Currie has pointed out: "In Quebec, rates must be filed with the provincial transport board which may amend any rates to make them fair and reasonable, and must approve changes proposed by the truckers. *Only British Columbia, Saskatchewan, and Manitoba carefully control rates for intra-provincial movements of freight.*"[34] Given the Board's disposition not to restrain unduly the private contracting activities of carriers and shippers during the relevant time span, it has been concluded that Quebec should be reclassified as a "Non-Regulating" province. Hence, there are three "regulating" provinces and seven "non-regulating" provinces in the statistical analysis that follows.

---

[30] The outputs were in no way affected by the preassigned classifications; the disparity involving Quebec was observed in the residuals of the estimating equations, which were entirely independent of any advance determinations as to the regulatory status of particular provinces. (Tables of residuals are available on request from the author.)

[31] The adoption of "agreed charge" pricing by Quebec trucking firms has been acknowledged by A. W. Currie [12].

[32] See [23], p. 3.

[33] See [23], p. 13 ff.

3. Estimating the rate effects from regulation

■ The objective of this study has been to find the effects on charges from restraining competition among suppliers of motor freight transportation services. The effects we are seeking to measure may be defined as the additional charges on shippers for the movement of goods by truck resulting from the political decision to regulate such movements.

In transportation, price is ordinarily identified as the rate assessed by a carrier firm for moving goods one ton-mile. Here, we are concerned with the operations of aggregations of suppliers and the level of their rates paid by large clusters of population, so that attention is centered on average rates per ton-mile. Our analysis has been structured to test the relationship of transport rates to selected operating costs of motor carriers serving particular markets. Conceptually this has led to the formulation of a multi-dimensional price-cost function, inasmuch as calculations have been performed by means of multiple regression techniques. Elements of demand were excluded deliberately so there should be no conflict between the effects of demand and cost on the level of transport rates. The influence of demand on price has been recognized,[35] but it is expected that the effects apply uniformly to the observations used here.

The hypothesis is that, in Canadian provinces regulating intra-provincial motor haulage, the rates received by intercity trucking firms exceeded those of firms providing service in non-regulating provinces, with due regard for dissimilarities of costs. Three tests have been made of the Canadian data on this presumption. First, multiple regressions have been calculated with motor carriers' revenue per ton-mile representing the dependent variable and designated cost factors constituting the independent variables. Second, a "*t*" test has been made to indicate the probability that the observed difference in the means of the residuals in the multiple regression for the regulating and non-regulating provinces was attributable to chance. One deficiency of the "*t*" test is its assumption that the underlying distribution is normal; although in the present context it appeared reasonable to accept this assumption, the non-parametric χ^2 test for goodness of fit was used to account for differences, given the possibility that the distribution was not normal.

Conceivably, the hypothesis that regulation of trucking resulted in higher revenues per ton-mile might be tested by the simple expedient of comparing the means of such statistics for aggregations of firms operating in regulating and non-regulating provinces, respectively. However, since revenues per ton-mile depended on other factors as well, and since these factors were subject to variations in each province, results obtained from such a procedure would be inconclusive. The effects of some of these factors were removed by means of the first stage of the model, based on the expression

---

[34] See [12], p. 459 (Emphasis supplied.)

[35] See, for example, Samuelson [24], p. 357 ff. An independent study of demand has been made, and it was found that for the Canadian trucking industry as a whole (excluding carriers domiciled in Newfoundland and Prince Edward Island), negative price elasticity, positive income elasticity, and positive cross elasticity with railroad transport prevailed. Cf. Sloss [25]. Computed elasticities of demand were: Price, − 1.044, Income, + 0.966, Cross (with railroads), + 1.055. All values were significant at the 0.95 level.

$$y^j = a + \sum_{i=1}^{n} b_i x_i^j + u^j, \qquad (1)$$

which can also be written as

$$y^j - a - \sum_{i=1}^{n} b_i x_i^j = u^j \qquad (2)$$

where y^j is a rate variable, x_i^j are cost variables, and u^j the residual differentials. In the second form, some of the rate variations due to the cost factors x_i^j have been subtracted from the observed rates y^j. Therefore, if it can be postulated that the cost variables represented by x_i^j have not been, themselves, affected by regulation, it follows that the differential effects of regulation on transport rates among the several provinces have been captured in u^j. In more technical terms, by comparing the values of the residuals, u^j, rather than the values of the dependent variables, y^j, the test of the hypothesis has not only been made more "efficient"[36] but the chances of obtaining spurious results due to the effects of the variables, x_i^j, have been greatly reduced.[37]

After some preliminary data adjustments, the first step of the model was fitted in the following algebraic form:

$$u^j = y^j - a - b_1x_1^j - b_2x_2^j - b_3x_3^j - b_4x_4^j - b_5x_5^j, \qquad (3)$$

where for each observation j, y^j was revenue or rate per ton-mile, and the independent variables were, in order: x_1^j average length of haul, x_2^j average net weight per loaded vehicle, x_3^j average fuel tax per gallon, x_4^j average license cost per truck or tractor per year, and x_5^j average annual wage per employee.

The second step of the model consisted of measuring the level of significance of the difference between the means of the residuals computed for the regulating and non-regulating provinces. A "t" test has been employed which, in effect, tested the "null" hypothesis that regulation does not affect the values of the residuals where

$$t = \frac{\bar{U}_r - \bar{U}_f}{\hat{\sigma}_{\bar{u}_r - \bar{u}_f}}, \qquad (4)$$

with $\bar{U}_r$ = the mean of residuals for regulating provinces, $\bar{U}_f$ = the mean of residuals for non-regulating provinces, and $\hat{\sigma}_{\bar{u}_r - \bar{u}_f}$ = the standard error of the difference between the means of the residuals.

---

[36] For a definition of the term "efficient" as used in this context, see Fraser [16], pp. 215–16.

[37] See Appendix 2 for an elaboration of computational procedure.

The final step in implementing the model involved the use of a χ^2 test for reasons already explained. This test statistic took the form

$$\chi^2 = \sum \frac{(f - f_c)^2}{f_c}, \tag{5}$$

with f representing the observed frequency, and f_c the computed frequency of the residuals arranged in a "2 × 2" table according to whether the signs of the residuals were positive or negative and whether the residuals applied to "regulating" or to "non-regulating" provinces.[38]

☐ **Testing the effects of regulation on intraprovincial trucking.** For truck traffic moving exclusively within the confines of a single province, the estimating model produced values for a constant and for the coefficients of the independent variables as shown in equation (6).

$$\begin{aligned} \hat{y} = 9.7691 &- \underset{(0.0056)}{0.0102x_1} - \underset{(0.1807)}{0.7779x_2} + \underset{(0.0740)}{0.1983x_3} \\ &- \underset{(0.0029)}{0.0096x_4} + \underset{(0.0007)}{0.0021x_5}. \end{aligned} \tag{6}$$

The standard error of the estimate (σ) = 1.1667. The coefficient of multiple determination (R^2) = 0.7241. The residuals for each observation were allocated to either the regulated or non-regulated category; excluding Newfoundland and Prince Edward Island from the model resulted in 18 residuals being assigned to the regulated category and 30 to the non-regulated category,[39] as shown in Table 6.

TABLE 6

THE RESIDUALS FROM INTRAPROVINCIAL REGRESSION

DESIGNATION OF CATEGORY	REGULATED	NON–REGULATED
ASSIGNED SUBSCRIPT	r	f
NUMBER OF OBSERVATIONS	$N_r = 18$	$N_f = 30$
TOTAL NUMBER OF OBSERVATIONS	$N = (N_r + N_f) = 48$	
DEGREES OF FREEDOM	$n = (N - 2) = 46$	
SUM OF RESIDUALS	$\Sigma U_r = +9.7568$	$\Sigma U_f = -9.7568$
MEAN OF RESIDUALS	$\bar{U}_r = +0.5420$	$\bar{U}_f = -0.3252$
STANDARD ERROR OF THE DIFFERENCE BETWEEN THE MEANS OF THE RESIDUALS (40)	$\hat{\sigma}_{\bar{u}_r - \bar{u}_f} = 0.3066$	

---

[38] See Croxton and Cowden [11], pp. 681–86, and Von Mises [34], p. 447 ff.

[39] The regulated category comprised Manitoba, Saskatchewan, and British Columbia. The non-regulated group included Nova Soctia, New Brunswick, Quebec, Ontario, and Alberta.

(The matrix of simple correlation coefficients appears in the Appendix in Table 17.)

These resulted in a value of

$$t = \frac{0.5420 + 0.3252}{0.3066} = 2.828\,. \tag{7}$$

This value in a one-tailed test with $n = 46$ yielded a (P) value below the 0.005-percent level of significance.[41] Hence, the null hypothesis of no categorical difference was rejected and the alternative assumption accepted that the mean excess of seven-eighths of a cent per ton-mile received by trucking firms operating under economic regulation was explainable as a consequence of regulation.

The hypothetical effects were confirmed by the Pearson's χ^2 test on the signs of the residuals. The following distribution shows these residuals,

observed signs of residuals	+	−	total
regulated	12	6	18
non-regulated	9	21	30
total	21	27	48

They produced a χ^2 value of 6.459, statistically significantly different from the hypothesized zero at the 2% level. This pattern of residuals is assumed to have followed from significantly higher prices in the regulated provinces.[42]

☐ **Testing the effects of regulation on extraprovincial and international trucking.** Data on movements across provincial borders or across the United States—Canadian boundary were analyzed the same way as in the foregoing section. The regression equation is

$$\begin{aligned} \hat{y}' = 8.8030 &- \underset{(0.0004)}{0.0014x_1'} - \underset{(0.0860)}{0.3838x_2'} + \underset{(0.0373)}{0.0782x_3'} \\ &- \underset{(0.0014)}{0.0057x_4'} + \underset{(0.0004)}{0.0009x_5'}\,. \end{aligned} \tag{8}$$

---

[40] The following procedure was followed in calculating the standard error.

$$\Sigma u_r^2 = \Sigma U_r^2 - \frac{(\Sigma U_r)^2}{N_r} \qquad \Sigma u_r^2 = 19.93 - \frac{95.20}{18} = 14.64$$

$$\Sigma u_f^2 = \Sigma U_f^2 - \frac{(\Sigma U_f)^2}{N_f} \qquad \Sigma u_f^2 = 37\ 24 - \frac{95\ 20}{30} = 34.07$$

$$\hat{\sigma}_{r+f} = \sqrt{\frac{\Sigma u_r^2 + \Sigma u_f^2}{N_r - 1 + N_f - 1}} \qquad \hat{\sigma}_{r+f} = \sqrt{\frac{14.64 + 34.07}{18 - 1 + 30 - 1}} = \sqrt{\frac{48.71}{46}} = \sqrt{1.0588}$$

$$= 1.0287$$

$$\hat{\sigma}_{\bar{u}_r - \bar{u}_f} = \sigma_{r+f}\sqrt{\frac{1}{N_r} + \frac{1}{N_f}} \qquad \hat{\sigma}_{\bar{u}_r - \bar{u}_f} = 1.0287\sqrt{\frac{1}{18} + \frac{1}{30}} = (1.0287)(0.298)$$

$$= 0\ 3066$$

The same procedure was followed in calculating similar values throughout the text. See Croxton and Cowden [11], p. 651 ff.

[41] See Edwards [14], p. 220.

[42] It should be noted that 33 of the 48 residuals displayed positive or negative signs conforming to *a priori* expectations. It could, accordingly, be inferred that 69 percent of the residuals showed "correct" signs, which would constitute additional evidence of a meaningful relationship between the existence or absence of provincial truck regulation, on the one hand, and the economic cost borne by users of intraprovincial motor freight service, on the other. See Croxton and Cowden [11], Appendix J, pp. 752–53.

TABLE 7

THE RESIDUALS FROM EXTRAPROVINCIAL REGRESSION

DESIGNATION OF CATEGORY	REGULATED	NON-REGULATED
ASSIGNED SUBSCRIPT	r'	f'
NUMBER OF OBSERVATIONS	$N_{r'} = 18$	$N_{f'} = 30$
TOTAL NUMBER OF OBSERVATIONS	$N = (N_{r'} + N_{f'}) = 48$	
DEGREES OF FREEDOM	$n = (N - 2) = 46$	
SUM OF RESIDUALS	$\Sigma U_{r'} = +1\ 2003$	$\Sigma U_{f'} = -1.2003$
MEAN OF RESIDUALS	$\bar{U}_{r'} = +\ 0.06668$	$\bar{U}_{f'} = -\ 0.04001$
STANDARD ERROR OF THE DIFFERENCE BETWEEN THE MEANS OF THE RESIDUALS	$\hat{\sigma}_{\bar{u}_{r'} - \bar{u}_{f'}} = 0.1651$	

The standard error of the estimate (σ) = 0.5832. The coefficient of multiple determination (R^2) = 0.7404. (The matrix of simple correlation coefficients is given in Appendix Table 18.)

The residuals were distributed between regulated and non-regulated categories based on extraprovincial truck regulation existing in the individual provinces from which transport originates. Details of this distribution have been provided in Table 7.

A "t" test to measure the level of significance of the difference of the sample means, $\bar{U}_{r'}$ and $\bar{U}_{f'}$, was calculated in conformity with equation (4), with $t = 0.646$. This value of "t" in a one-tailed test has a significance level (P) between 0.25 and 0.30,[43] so that there is no basis for concluding that per ton-mile revenues received by Canadian trucking firms on their extraprovincial and international services have been increased by entry or price controls.[44] Similarly the outcome of the χ^2 test, with one degree of freedom, indicated a level of significance (P) between 0.20 and 0.25.[45] Economic regulation could not be held accountable for higher rates to shippers of goods across some provincial borders.

☐ **Testing the effects of regulating all Canadian trucking services.** The same computational procedures were used in connection with data

---

[43] See [11], Appendix I, pp 750–51.

[44] Sir Ronald A. Fisher [15] commented as follows on the outcome of a "t" test "The value of P is, therefore, between 0.1 and 0 05 and cannot be regarded as significant." p 126.

[45] Actual values of the 48 residuals included 28 residuals with positive or negative signs consistent with *a priori* assumptions. Hence, only 58 percent of the signs of the residuals could be considered "correct" in contrast to the 69 percent so indicated for the intraprovincial trucking analysis. The Pearson's χ^2 test on the basis of the distribution of the residuals is as follows

observed signs of residuals	+	−	total
regulated	11	7	18
non-regulated	13	17	30
total	24	24	48

$$\chi^2 = \frac{(11-9)^2}{9} + \frac{(13-15)^2}{15} + \frac{(7-9)^2}{9} + \frac{(17-15)^2}{15} = 1.422\,.$$

This results in a level of significance (P) between 0.20 and 0.25. Cf. Croxton and Cowden, [11], Appendix J.

TABLE 8

THE RESIDUALS FROM INTRAPROVINCIAL AND EXTRAPROVINCIAL REGRESSION

DESIGNATION OF CATEGORY	REGULATED	NON-REGULATED
ASSIGNED SUBSCRIPT	$r^{\#}$	$f^{\#}$
NUMBER OF OBSERVATIONS	$N_r\# = 18$	$N_f\# = 30$
TOTAL NUMBER OF OBSERVATIONS	$N = (N_r\# + N_f\#) = 48$	
DEGREES OF FREEDOM	$n = (N - 2) = 46$	
SUM OF RESIDUALS	$\Sigma U_r\# = +7.6829$	$\Sigma U_f\# = -7.6829$
MEAN OF RESIDUALS	$\bar{U}_r\# = +0.4268$	$\bar{U}_f\# = -0.2561$
STANDARD ERROR OF THE DIFFERENCE BETWEEN THE MEANS OF THE RESIDUALS	$\hat{\sigma}_{\bar{u}_r\# - \bar{u}_f\#} = 0.2581$	

covering all intraprovincial, extraprovincial and international trucking services performed by Canadian firms. Expressed in the form of equation (1), with the dependent variable representing rates per ton-mile and the independent variables denoting, as before, average length of haul, average load per vehicle per trip, average fuel tax per gallon, average license cost per truck or tractor, and average wage per employee, in that order, the estimating equation is

$$\begin{aligned} \hat{y}^{\#} = 10.7241 &- \underset{(0.0028)}{0.0036x_1^{\#}} - \underset{(0.1473)}{0.7787x_2^{\#}} + \underset{(0.0603)}{0.1642x_3^{\#}} \\ &- \underset{(0.0024)}{0.0069x_4^{\#}} + \underset{(0.0007)}{0.0015x_5^{\#}}\,. \qquad (9) \end{aligned}$$

The standard error of estimate (σ) = 0.9729. The coefficient of multiple determination (R^2) = 0.8117. (A correlation matrix for this equation has been tabulated in Appendix Table 19.)

The allocation of residuals between the regulating and non-regulating provinces was completed for the consolidated data as had been done for the residuals produced by the estimating equations covering intraprovincial trucking and extraprovincial-international trucking, respectively. The results are shown in Table 8.

The "t" test for the significance of the difference of the two sample means, $\bar{U}_r^{\#}$ and $\bar{U}_f^{\#}$, shows $t = 2.646$, which in a one-tailed test with $n = 46$ indicated a level of significance (P) between 0.005 and 0.025.[46] This supports the conclusion that rates per ton-mile received by Canadian motor carriers were materially increased when and where there was regulation by provincial transport commissions. Even with the "averaging" effect from pooling the two samples, there

---

[46] See Edwards, [14], Table IV, p. 220.

is indication of a rate-increasing tendency in intraprovincial regulation.[47]

☐ **Estimating the cost of regulation in Canada.** The model shows probable rate increases attributable to the imposition of controls over important segments of the Canadian trucking industry. For intraprovincial truck traffic, a difference between the means of the residuals in the regulating and non-regulating provinces was approximately 0.8672 cents per ton-mile. This difference equals the additional revenue per unit of service received by trucking firms transporting within a regulating province over the ton-mile revenue obtained by firms operating in non-regulating provinces. This rate difference multiplied by ton-mileage shows how much more those buying the regulated service have to pay for regulation. The total annual increment of charges in terms of additional payments by users of truck transportation has been estimated at slightly less than $10 million, as shown in Part I of Table 9.

For all commercial truck traffic moving intraprovincially, extraprovincially, and between Canada and the United States in trucks of Canadian registry, a difference between the means of the residuals in the regulating and non-regulating provinces of 0.6829 cents per ton-mile has been indicated in Table 8. Then for all transportation the estimated annual buyers' costs from regulation have been approximately $16 million (on the basis of calculations shown in Part II of Table 9). There would seem to have been substantially higher payments for trucking service in areas of Canada where regulation prevailed during the 1958–1963 period.

☐ **The comparable costs of regulating trucking in the United States.** The lack of quantitative information on the activites of unregulated trucking companies in the United States prohibits calculation of regressions such as those in the last section. But it is possible to develop some conjectural approximations of the added charges on users of trucking service from regulation in this country.

To employ the regressions for this exercise, data was assembled relating to regulated common carriers of Classes I and II in the United States during the six years from 1958 through 1963.[48] This information was derived from the Interstate Commerce Commission's annual statistical series *Transport Statistics in the United States*, part 7 of which applies to motor carriers. Conforming to the Commission's practice, the data for each year were broken down among nine geographic regions; these data have been consolidated for Classes I

---

[47] For completeness here, it should be noted that 33 of the 48 residuals manifested positive or negative signs compatible with *a priori* expectations. Hence, the same number of "correct" signs were obtained as in the case of the intraprovincial analysis. Applying a χ^2 test to these residuals developed the following results

observed signs of residuals	+	−	total
regulated	13	5	18
non-regulated	10	20	30
total	23	25	48

With $\chi^2 = 6.82$.

The figure of 6.82 with one degree of freedom has a value for (P) between 0.005 and 0 01, which would be deemed statistically significant. Cf Croxton and Cowden [11], Appendix J

[49] Throughout these years, the Interstate Commerce Commission defined Class I carriers as those trucking firms reporting annual gross revenues of $1 million or more, Class II carriers report annual gross revenues between $200,000 and $1 million, Class III carriers report annual gross revenues less than $200,000.

TABLE 9

ESTIMATING THE ANNUAL COST OF APPLYING ECONOMIC REGULATION TO INTERCITY COMMERCIAL TRUCKING OPERATORS IN CANADA

(I) MANITOBA, SASKATCHEWAN, BRITISH COLUMBIA. INTRAPROVINCIAL TRAFFIC ONLY

YEAR	MILLION NET TON–MILES (FROM D.B.S. CAT. NO. 53-207)	ESTIMATED COST OF REGULATION @ 0.8672¢ PER TON–MILE (FROM EQUATIONS (4) AND (6))
1958	813	$7,050,336
1959	933	8,090,976
1960	940	8,151,680
1961	1,269	11,004,768
1962	1,389	12,045,408
1963	1,451	12,583,072
TOTAL	6,795	$58,926,240
MEAN	1,133	$ 9,821,040

(II) MANITOBA, SASKATCHEWAN, BRITISH COLUMBIA: ALL INTERCITY TRAFFIC PERFORMED BY VEHICLES REGISTERED IN THESE PROVINCES

YEAR	MILLION NET TON–MILES (FROM D.B.S. CAT. NO. 53-207)	ESTIMATED COST OF REGULATION @ 0.6829¢ PER TON–MILE (FROM EQUATIONS (12) AND (13))
1958	1,494	$10,202,526
1959	1,786	12,196,594
1960	1,928	13,166,312
1961	2,525	17,243,225
1962	3,023	20,644,067
1963	3,266	22,303,514
TOTAL	14,022	$95,756,238
MEAN	2,337	$15,959,373

and II and represent the operations of carriers responsible for approximately 70 percent of total revenues realized by all motor freight carriers filing reports with the Interstate Commerce Commission.[49]

The first step in using the data was to arrange the U. S. material so it could be compared, and later combined, with the Canadian figures utilized above. Since no suitable information respecting fuel taxes per gallon for United States carriers was available, the variable for this item (x_8) was omitted from both the U. S. and succeeding Canadian equations. The dependent variable and the balance of the independent variables were the same as those employed in the regressions above.

---

[49] The nine geographic regions included New England, Middle Atlantic, Central, Southern, Northwestern, Midwestern, Southwestern, Rocky Mountain, and Pacific.

U. S. motor carriers represented in this statistical analysis accounted for the following percentages of total revenues received by all trucking firms filing reports with the Interstate Commerce Commission. 1958, 70.4%; 1959, 73.0%; 1960, 68.9%; 1961, 71.5%; 1962, 72.1%; 1963, not available. "Concentration in motor carrier of property industry 1957–62," in [31], p. 149, Table 33.

The estimating equation covering commercial trucking in the U. S. is

$$\hat{y}^* = 20.0607 - \underset{(0.0021)}{0.0166x_1^*} - \underset{(0.1667)}{1.4081x_2^*} + \underset{(0.0006)}{0.0032x_4^*} + \underset{(0.0002)}{0.0004x_5^*}\,, \qquad (10)$$

while the comparable estimating equation for all Canadian for-hire trucking (recomputed to exclude the variable (x_3)) is

$$\hat{y}^{\#\#} = 14.8294 - \underset{(0.0027)}{0.0067x_1^{\#\#}} - \underset{(0.1471)}{0.6331x_2^{\#\#}} - \underset{(0.0025)}{0.0058x_4^{\#\#}} + \underset{(0.0006)}{0.0006x_5^{\#\#}}\,. \qquad (11)$$

A moderately close resemblance between the signs and coefficient values in the two equations is obvious; only the signs for the variable (x_4), average annual license cost, were in disagreement. Each coefficient in both equations had a value at least twice as great as its standard error with the exception of that associated with ($x_5^{\#\#}$), average annual wage per employee, in the second equation. Considering differences in area, climate, population densities and dispersions, industrial locations and product mix, as well as other relevant factors, it would have been unrealistic to anticipate more correspondence between the United States and Canadian trucking industry. Close examination of the correlation matrices in Appendix Tables 20 and 21 yields explanations for some of the more noticeable disparities; among these would be the fact that average length of haul (x_1) ranked first among the independent variables in terms of its contribution to the coefficient of multiple determination (R^2) in the United States equation while average weight per load (x_2) ranked second, whereas in the Canadian equation these rankings were reversed. However, as seen in Table 10, the average length of haul was almost twice as long for U. S. trucking firms as for those in Canada, although the average net weight per loaded vehicle was virtually the same for both countries.[50]

The regressions seem sufficiently similar to allow the second step in estimating the cost of regulation in the United States trucking industry. The data for the American and Canadian equations have been combined into a single set, which was used to calculate the following regression

$$\hat{y}'' = 14.7933 - \underset{(0.0015)}{0.0095x_1''} - \underset{(0.0808)}{0.8221x_2''} + \underset{(0.0005)}{0.0004x_4''} + \underset{(0.0001)}{0.0004x_5''}\,. \qquad (12)$$

The standard error of the estimate (σ) = 1.0374. The coefficient of multiple determination (R^2) = 0.7406. (The matrix of simple correlation coefficients is given in Appendix Table 22.)

---

[50] Additional information comparing equations (10) and (11) follows:

	Equation (10) U S.A.	Equation (11) Canada
The standard error of the estimate (σ)	0.6936	1.0430
The coefficient of multiple determination (R^2)	0.8752	0.7784

Correlation matrix information appears in Appendix 3, Tables 20 and 21.

TABLE 10

A COMPARISON OF STATISTICS FOR UNITED STATES AND CANADIAN FOR-HIRE TRUCKING FIRMS COVERING THE YEARS 1958 THROUGH 1963

DESCRIPTION OF VARIABLE	SYMBOL AND ESTIMATING EQUATION NO.		COUNTRY	MEAN VALUE OF VARIABLE	STANDARD DEVIATION OF THE MEAN	CONTRIBUTION TO R^2	
						%	RANK
REVENUE PER TON-MILE	y^*	(10)	UNITED STATES	6.6537 CENTS	(1.8879)		
	$y^{\#\#}$	(11)	CANADA	6.9688 CENTS	(2.1191)		
AVERAGE LENGTH OF HAUL	x_1^*	(10)	UNITED STATES	234.17 MILES	(56.44)	0.6595	1
	$x_1^{\#\#}$	(11)	CANADA	121.46 MILES	(88.65)	0.0860	2
AVERAGE NET WEIGHT PER LOAD	x_2^*	(10)	UNITED STATES	10.73 TONS	(1.46)	0.1232	2
	$x_2^{\#\#}$	(11)	CANADA	10.20 TONS	(1.97)	0.6575	1
AVERAGE ANNUAL LICENSE COST PER TRUCK OR TRACTOR	x_4^*	(10)	UNITED STATES	\$882.52	(374.59)	0.0810	3
	$x_4^{\#\#}$	(11)	CANADA	\$487.17	(116.25)	0.0302	3
AVERAGE ANNUAL WAGE PER EMPLOYEE	x_5^*	(10)	UNITED STATES	\$6,446.43	(541.94)	0.0115	4
	$x_5^{\#\#}$	(11)	CANADA	\$3,928.54	(611.14)	0.0046	4

SOURCES: UNITED STATES STATISTICS DERIVED FROM COMPUTER PRINT-OUT FOR EQUATION (10).
CANADIAN STATISTICS DERIVED FROM COMPUTER PRINT-OUT FOR EQUATION (11).

An allocation of residuals between regulating and non-regulating areas was made on the basis that the five unregulated provinces (Nova Scotia, New Brunswick, Quebec, Ontario, and Alberta) comprised the unregulated sector, while the remaining three Canadian provinces and the entire nine geographic regions of the United States were the regulated sector.[51] Of the total 102 observations, 30 fell into the non-regulating and 72 into the regulating categories. Results are given in Table 11.

The value of "*t*" for the difference of the two sample means in Table 11 has been computed to be 2.0597, indicating a level of significance (*P*) between 0.005 and 0.025.[52] It can be concluded that revenues per ton-mile received by United States and Canadian regulated freight haulers were 0.45 cents per ton-mile greater than those of Canadian unregulated trucking firms during the 1958–1963 period.[53]

---

[51] All U. S. data, having been derived from Interstate Commerce Commission sources, were assumed to cover regulated truck transportation even though some small and undeterminable portion included therein may have consisted of commodities exempted from interstate regulation by sec. 203(b)(6), Part II of the *Interstate Commerce Act*.

[52] See Edwards [14], p. 220, Table IV

[53] Only 57 of the 102 residuals exhibited positive or negative signs conforming with *a priori* anticipations. A χ^2 test based on the distribution of these residuals developed the following results:

observed signs of residuals	+	−	total
regulated	35	37	72
non-regulated	8	22	30
total	43	59	102

With $\chi^2 = 4.187$

This value has a (*P*) value between 0.025 and 0.05, which although not as conclusive as the outcome of the "*t*" test, is nonetheless within the range of statistical significance. Cf. Croxton and Cowden [11], Appendix J.

TABLE 11

THE RESIDUALS FROM REGRESSION OF COMBINED U.S AND CANADIAN DATA

DESIGNATION OF CATEGORY	REGULATED	NON-REGULATED
ASSIGNED SUBSCRIPT	r''	f''
NUMBER OF OBSERVATIONS	$N_{r''} = 72$	$N_{f''} = 30$
TOTAL NUMBER OF OBSERVATIONS	$N = (N_{r''} + N_{f''}) = 102$	
DEGREES OF FREEDOM	$n = (N - 2) = 100$	
SUM OF RESIDUALS	$\Sigma U_{r''} = +9.4839$	$\Sigma U_{f''} = -9.4839$
MEAN OF RESIDUALS	$\bar{U}_{r''} = +0.1317$	$\bar{U}_{f''} = -0.3161$
STANDARD ERROR OF THE DIFFERENCE BETWEEN THE MEANS OF THE RESIDUALS	$\hat{\sigma}_{u_{r''}} - \bar{u}_{f''} = 0.21741$	

The magnitude of the charges on users has been estimated two ways. First, the mean revenue per ton-mile for U. S. Class I and II common carriers for the years 1958 through 1963 was shown in Table 10 to equal 6.6537 cents. The estimated increased revenue per ton-mile for these carriers as a result of regulation was found to be 0.4478 cents. Had regulation not been applied to U. S. common carriers, it may be surmised that their revenues per ton-mile would have been reduced by this 6.73 percent. This would be equivalent to a transportation saving of charges on users of $348 million per year, arrived at by applying the 6.73 percent figure to average annual receipts of $5.177 billion earned by these firms between 1958 and 1963.

A second method of estimating the cost of economic regulation of motor haulage in the United States has been set forth in Table 12. This approach indicates revenue increases to truckers of $361 million per year ascribable to regulation. No claims of precision are implied for either of the estimates. Suffice it to point out that the approaches were dictated by problems related to the availability of U. S. data.

4. The effects of regulation on profitability

■ Various measures of profitability might have been incorporated in the model as an additional independent variable. They were not for two main reasons: (1) The amount of profit, reported as net income or as net operating revenue, would result from rate increases rather than be their cause. (2) Computations executed for the purpose of investigating possible relationships between profitability and other variables in the estimating equations proved inconclusive. For example, the ratio of net operating revenue to gross operating revenue, a standard of profitability uniformly applied in the trucking industry,[54] was correlated with the dependent variable, $y^{\#}$, in equa-

---

[54] On this point, see Phillips [21], pp. 271 and 275.

TABLE 12

ESTIMATING THE ANNUAL COST OF APPLYING ECONOMIC REGULATION TO INTERCITY COMMERCIAL TRUCKING OPERATIONS IN THE UNITED STATES

ALL UNITED STATES CLASS I AND II INTERCITY COMMON CARRIERS OF PROPERTY		
YEAR	MILLION NET TON–MILES (FROM I.C C. *TRANSPORT STATISTICS*)	ESTIMATED COST OF REGULATION @ 0.4478¢ PER TON–MILE (FROM EQUATIONS (18) AND (19)
1958	66,446	$297,545,188
1959	78,677	352,315,606
1960	75,144	336,494,832
1961	82,638	370,052,964
1962	86,828	388,815,784
1963	94,013	420,990,214
TOTAL	483,746	$2,166,214,588
MEAN	80,624	$ 361,035,764

tion (9) above. A very low coefficient of correlation (r) (+ 0.0986) resulted. The same indicator of profitability regressed against the residuals for the regulated and non-regulated provinces derived from equation (9) produced coefficients of correlation (r) of − 0.281 and − 0.126, which gave no evidence of statistical significance. Finally, the ratio of net to gross operating revenue experienced by all Canadian for-hire intercity trucking firms, collectively, from 1958 through 1963, equaled 6.42 percent in the case of firms based in the non-regulating provinces, including Newfoundland and Prince Edward Island, compared with a ratio of 6.12 percent for firms having their headquarters in the regulating provinces.[55] This outcome was somewhat paradoxical in view of the indications above that carriers in the non-regulating provinces received average revenues per ton-mile 0.68 cent below the level found for trucking companies in the regulating provinces.

5. Economies of scale and concentration of firms

■ The Chief of the Transportation Section, Dominion Bureau of Statistics, has pointed out that "Trucking has developed as a type of transportation characterized by many small fleets, quickly adaptable to changes, qualitative or quantitative, according to requirements . . . . The foregoing is not to suggest that there are no economies of scale (that is advantages of size) in trucking . . . . The real point worth noting here is that a substantial proportion of the overall requirement for truck transportation is so varied in type and location that it has been, and will probably continue to be, best met by small business units." On the subject of concentration, *per se*, in Canadian for-hire trucking, the same individual has noted: "While there are a number of large firms in the 'For Hire' sector, none is sufficiently big to suggest the existence of anything other than a vigorously competitive market. Indeed, the top ten firms, in terms of revenue, account for only one-sixth of total revenue, whilst the top fifty earn a third."[56] This conclusion may be important for explaining rate-cost differences between provinces.

---

[55] See [4] and [7]. It was necessary to include Newfoundland and Prince Edward Island in these calculations because the above cited statistical sources provided data for the four Atlantic provinces in aggregated form only. Both provinces were assigned to the "non-regulating" classification in compliance with the assignment schedule in the text.

One possible influence of economies of scale on the values of the residuals in the estimating equations merited careful scrutiny. This was investigated using the reports published by the Canadian Dominion Bureau of Statistics, which divided trucking firms into four classifications according to annual gross revenue. The presence of economies of scale could be detected by a superior profit realization for the group comprising the largest firms in comparison with the earnings of the industry as a whole.[57]

The number of firms in the group made up of the largest operators accounted for 3.1 percent of the total for the seven non-regulating provinces and 2.5 percent of the total in the three regulating provinces. As shown in Table 13, the percentage of annual *gross* operating revenues accruing to the firms in the largest group was 65.5 percent and 50.7 percent, respectively, for the non-regulating and regulating classifications; corresponding ratios of *net* operating revenues were 44.9 percent and 19.3 percent. A measure of the influence of economies of scale was obtained from the expression

$$S = \frac{\sum \text{N.O.R.: GRP. I}}{\sum \text{N.O.R.: All GRPS.}} \Big/ \frac{\sum \text{G.O.R.: GRP. I}}{\sum \text{G.O.R.: ALL GRPS.}} .$$

The Profitability Ratio ascribable to size of firm = S.
Net operating revenue = N.O.R.
Gross operating revenue = G.O.R.
Firms having annual gross revenue of $500,000 or more = GRP. I.
The totality of all firms in all groups = ALL GRPS.

A value for "S" exceeding unity would suggest that the profits of the larger firms were disproportionately high compared with the earnings of the entire industry; a value for "S" below unity should lead to the opposite conclusion. In the present case the composite value of "S" was 0.381 for the regulating provinces and 0.686 for the non-regulating provinces, as shown in Table 13. In no single province was the ratio equal to or exceeding unity. This confirms the view generally held concerning the absence of important economies of scale in trucking.[58] Although the largest firms represented a higher percentage of the total number in the non-regulating provinces than in the regulating provinces and, in addition, displayed higher profitability ratios, their actual performance does not suggest that there were substantial economies of scale.

6. Self-regulation by the Canadian trucking industry

■ An important problem relating to the validity of those findings concerns the influence on motor carrier rate levels exerted by various rate bureaus and tariff publishing agents representing member firms

---

[56] See Steeves [27], p. 3.

[57] The Dominion Bureau of Statistics in [4] separated operating and financial statistics for common and contract carriers, listed in segregated sections, according to annual gross revenues. Group I consisted of carriers having annual gross revenues of $500,000 or more; group II included firms with gross revenues from $100,000 to $499,999 per year, group III firms reported annual gross revenues from $20,000 to $99,999; and group IV contained operators grossing less than $20,000 annually. Each carrier's statistics were shown in the province where its head office was located.

[58] This subject has been comprehensively treated in [3], Chapter 1.

TABLE 13

THE RATIO OF NET OPERATING REVENUES OF GROUP I TRUCKING FIRMS TO NET OPERATING REVENUES OF ALL TRUCKING FIRMS

DIVIDED BY

THE RATIO OF GROSS OPERATING REVENUES OF GROUP I TRUCKING FIRMS TO GROSS OPERATING REVENUES OF ALL TRUCKING FIRMS

PROVINCE	REGULATORY STATUS	NUMERATOR	DENOMINATOR	QUOTIENT
ATLANTIC*	(N.R.)	6.4	22.2	0.288
QUEBEC	(N.R.)	45.0	63.5	0.710
ONTARIO	(N.R.)	50.3	69.4	0.725
MANITOBA	(R.)	36.2	71.5	0.505
SASKATCHEWAN	(R.)	17.7	29.4	0.601
ALBERTA	(N.R.)	32.6	61.0	0.534
BRITISH COLUMBIA	(R.)	12.5	42.7	0.292
WEIGHTED AVERAGES	ALL (R.) COMBINED	19.3	50.7	0.381
	ALL (N.R.) COMBINED	44.9	65.5	0.686

* ATLANTIC PROVINCES INCLUDE NEWFOUNDLAND, PRINCE EDWARD ISLAND, NOVA SCOTIA AND NEW BRUNSWICK.

(R.) DESIGNATES PROVINCES APPLYING ECONOMIC REGULATION TO INTERNAL FOR–HIRE TRUCKING.

(N.R.) DESIGNATES PROVINCES NOT REGULATING ECONOMIC ASPECTS OF INTERNAL FOR–HIRE TRUCKING.

SOURCE OF DATA: DOMINION BUREAU OF STATISTICS–REPORTS BEARING CATALOGUE NUMBERS 53-205, (1958–1960); 53-221, (1960–1963); 53-222 AND 53-223, (1961–1963).

in the Canadian trucking industry. These organizations, constituting a form of industry-sponsored regulatory system, were conceptually in a position to enforce a measure of control over prices of their own. This control may or may not have replaced or enhanced the effects of regulation.

There were in 1960 three major tariff bureaus in Canada acting on behalf of truck lines operating in British Columbia, Ontario, Quebec, and the Maritime Provinces. Carriers participating in one of these bureaus did not ordinarily publish their own tariffs but relied entirely on bureau publications. These were filed with provincial boards and considered as the legal regulated rates, except in Ontario where there was no rate control.[59] In 1960, also, it was declared not legal for carriers in Alberta to combine in a tariff bureau and issue an agency tariff; such action would have made them subject to *The Combines* (antitrust) *Act.*[60]

It was reported in a 1962 Royal Commission Study that in provinces which did not control rates the trucking industry "had developed its own techniques for coordination of rate making on such traffic as it was deemed advisable . . . few firms were prepared to bind themselves entirely to following a uniform tariff schedule. . . . "[61] Additional indications of the extent of tariff bureau authority appeared in the questionnaire responses from the provincial regulatory boards and in a province-by-province examination of the residuals produced by the regression equations. The accumulated evidence points to Ontario as the only province where rate bureau membership could have produced a measurable upward adjustment in rates. The situation prevailing in that province in 1961 is best described by a trucking company official: "I don't agree with secret [discount] rates as they were just called. We have them in Ontario as everyone knows. I cannot say that the tariffs as published by the Canadian Transport Tariffs [Bureau] are entirely paper rates. Possibly they are providing a clue for the secret rates, they are a guide."[62]

---

[59] Filing of intraprovincial truck rates became effective in Ontario on May 1, 1963

[60] See the remarks of Mr. Mikkelson, vice president-traffic, Canadian Freightways, Ltd. during a conference held on February 17, 1960, on "Private Transportation. Its Impact on For-Hire Carriers," [8], p. 142.

7. A final evaluation and summary

■ In judging the worth of the approach used here, it is necessary to assess the quality and reliability of the statistical data utilized in the many regression equations. Once again, we draw upon the views of the Chief of the Transportation Section of the Dominion Bureau of Statistics, Mr. E. T. Steeves, whose organization compiled and published most of the Canadian data employed in this inquiry: "All statistics derived from the Motor Transport Traffic Survey[63] are, of course, subject to the normal ranges of uncertainty associated with sample surveys. This range of uncertainty varies according to the type of data. . . . Survey results are also affected by such factors as poor reporting and poor response. For these reasons, caution should be exercised in using figures for individual years. In assessing trends over the period, for example, a simple comparison between 1957 and 1967 would be less reliable than some other method such as the fitting of a trend line."[64] Since the data in question were aggregated over entire regions and, moreover, were used essentially for comparative purposes, the problems referred to have been diminished, but caution is still in order.

The quantitative estimates of the effects of entry controls and price controls by government on a particular form of transportation have shown a pattern of results. Applied to the motor freight carriers in Canada, it appears that the charges of regulation on final users of transport services vary between approximately ten and fifteen million dollars per year, the former figure representing intraprovincial operations exclusively, the latter all traffic carried by Canadian motor vehicles. Both estimates have been based on statistics covering the six years commencing with 1958 and concluding with 1963. An indirect application of similar estimating procedures to motor common carriers of freight in the United States shows the charges from regulation ranging between $348 and $361 million per year in this period. We conclude that with a probability of over 95 percent, trucking services performed in regulated areas collected more revenues than those in non-regulated areas.

---

[61] See [9], pp. 37–38.

[62] See the comments of Mr. Graves, traffic manager, Motorways (Ontario), Ltd., Toronto, during a conference held February 15 and 16, 1961, on "Transportation–Its Role in Reducing Distribution Costs," in [8], p. 270.

[63] See [4], and [7].

[64] See Steeves [27], Statistical Appendix, p. 2.

In the light of the findings, it may be possible to envisage the amount of a nation's payments for trucking services being substantially reduced by rates established through the workings of non-regulated markets. We suggest that such a view be taken into account in the formulation of regulatory policies applicable to these services in both the United States and Canada.

Appendix 1

Comments of officials of provincial regulatory agencies and representatives of trucking organizations relating to the status of regulation of the for-hire motor freight industry in Canada[65]

Nova Scotia

Motor carriers operating in this Province are required to file their rates with [the Board of Commissioners of Public Utilities], and our *Motor Carrier Act* requires that carriers must carry freight at the filed rates. However, the approval of filed rates [by the Board] is not required in connection with freight tariffs.

Failure to carry at filed rates is a violation of our *Motor Carrier Act* resulting in prosecution where such fact is established; but [the Board] has received very few complaints with respect to any such violations.

New Brunswick

Regulation 49 made under the *Motor Carrier Act* requires that rates be filed with [The Motor Carrier Board] and the carrier must not charge any rate other than that approved by the Board. The Board does not have the facilities to check rates, but automatically accepts any rate filed by the carrier. . . . In cases where complaints have been received action has been taken against the carrier.

Regulation 49.(b) as amended August 1st, 1957, reads as follows: "No person being the operator of a public motor truck shall accept goods for carriage or carry goods at a rate of compensation in excess of the rate approved by the Board."

On September 27th, 1961, the regulation was amended to read as follows, and this Regulation is the one currently in effect: "No person being the operator of a public motor truck shall accept goods for carriage or carry goods *at any rate other than that* approved by the Board." (Change in wording italicized.)

All Atlantic Provinces (Nova Scotia, New Brunswick, Prince Edward Island, and Newfoundland)

From experience with the motor carrier boards and the industry, generally speaking, the carriers (both common and contract) are required to file their rates and to observe them. However, in actual practice every change in rates is not filed. It follows, therefore, that the carriers do depart from the rates filed in those instances.

Generally speaking carriers belonging to the bureaus adhere to the bureau publications more closely than the non-bureau members adhere to their own tariffs.

---

[65] These comments, all written in 1969, have been excerpted from personal communications to the writer, so their authorship has not been disclosed. All are available for examination in the writer's files.

The enforcement of filed or approved rates is largely limited to cases where complaints have been received. This does not imply that the carriers do not generally adhere to their filed rates—it does indicate that goods do, in fact, move at other than filed or approved rates and no comprehensive efforts are made to enforce the filed or approved rates.

Ontario

Sections 23 to 29 of Regulation 503, which became effective May 1, 1963, had the effect of rate filing only. The regulatory body, therefore, has no control over the level of rates assessed by the individual carriers, either through bureau membership or individually filed tariffs on behalf of carriers who do not wish to join such a bureau.

The immediate effect of rate filing was a general reduction in individual rates to the lowest denominator of those filed. Subsequent general increases have brought the levels to a more realistic basis but, as previously stated, inasmuch as the rates per se are not regulated, tolls are often reduced on volume movements, generally by what could be termed smaller carriers or those having a proportionately reduced overhead.

Manitoba

The requirement is that an applicant [for operating rights] must prove either that the present service is insufficient or that public convenience will be promoted. Note that only one, and not both, is required to be proven.

[The Motor Transport Board] has set up a standard Tariff of Tolls and Regulations and Conditions of Carriage covering the Intra-provincial trucking although the individual operator may file a rate other than the statutory one.

The statutory rates set up by the Board are required to be complied with Intra-provincially. No control is maintained by the Board over Extra-provincial rates.

The statutory rates are violated on occasions by carriers, and the Board attempts to police and investigates alleged Tariff violations.

International movements between Ontario and the United States

Basically, from the beginning of regulation of motor carriers in the United States, traffic handled by motor carrier has moved between points in the U. S. and points in Ontario on through one factor rates. Where traffic moves to points where there are no through rates published, the rates are computed over points in the United States . . . . The Interstate Commerce Commission has held consistently that where the traffic moves on through rates that they are subject to all provisions of Part II of the Act. They have even held further that a carrier operating wholly within Ontario must assess the published tariff rate if he is a party to it and if he elects not to do so, that the carriers holding Interstate Commerce authority will be required to eliminate their interchange with the Ontario carrier.

With regards to the regulations of the Ontario Highway Transport Board, Sections 23 to 29 have complete full application. Thus, rates filed between points in the United States and points in Ontario are subject both to the regulations of the Interstate Commerce Commission and to those of the Ontario Highway Transport Board. While this does create some problems, the cooperation of the two regulatory bodies is such [that] the difficulties are held to a bare minimum.

Appendix 2

Computational procedure

■ In making the calculations, data pertaining to one province for one year was treated as an "observation." Sources of these statistics included reports of the Dominion Bureau of Statistics, an agency of the Federal Government of Canada with headquarters in Ottawa.[66]

Adjustments were made after a number of preliminary least squares equations were calculated. The regressions took the form of equation (1). The experimental trials portended generally successful results, subject, however, to certain modifications. These involved eliminating all observations pertaining to the provinces of Newfoundland and Prince Edward Island; employing separate schedules of input data for intraprovincial services and extraprovincial-international services, respectively; and excluding certain items from the estimating equation because of a high incidence of multicollinearity among independent variables disclosed by the simple correlation matrices. The computational procedure adopted in light of these determinations entailed processing three sets of data covering (1) intraprovincial commercial intercity trucking movements, exclusively, (2) extraprovincial and international motor freight haulage, and (3) all intercity service rendered by all for-hire trucks of Canadian registry. All three data sets correlated the dependent variable, revenue per ton-mile, with the five independent variables designated in the text. Forty-eight observations were utilized in each of the three processings consisting of statistics for eight provinces during each of the six years from 1958 through 1963.

Newfoundland and Prince Edward Island were excluded from the model because of (1) the relatively small size of their for-hire truck populations, (2) their insular locations, resulting in unique transportation problems, (3) caution expressed in D.B.S. reports that estimates for these provinces contained a large degree of sampling error, and (4) the impracticability of prorating to these provinces a reasonably accurate share of their costs and revenues when published in consolidated form for the Atlantic Provinces as a whole.

---

[66] See [4], [5], and [7].

TABLE 14

PROVINCIAL REGULATION OF TRUCKS: 1959

	PRINCE EDWARD ISL.		NOVA SCOTIA		NEW BRUNSWICK		QUEBEC		ONTARIO		MANITOBA		SASKAT-CHEWAN		ALBERTA		BRITISH COLUMBIA	
GOVERNING MOTOR CARRIER ACTS	MOTOR CARRIER ACT		MOTOR CARRIER ACT		MOTOR CARRIER ACT		TRANSP'N. BOARD ACT		PUBLIC COM. VEHICLES ACT		HIGHWAY TRAFFIC ACT		THE VEHICLES ACT		PUB. SERV. VEHICLES ACT		MOTOR CARRIER ACT	
MOTOR CARRIER REGULATORY AUTHORITIES	PUBLIC UTIL. COMM.		BOARD OF COMM'RS. OF PUB. UT.		MOTOR CARRIER BOARD		TRANSP'N. BOARD		HIGHWAY TRANSPORT BOARD		MOTOR CARRIER BOARD		HIGHWAY TRAFFIC BOARD		HIGHWAY TRAFFIC BOARD		PUBLIC UTILITY COMMISSION	
ITEMS	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
REQUIREMENT TO SHOW PUBLIC CONV. & NECESSITY	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
CARRIER OPERATING LICENSES (FRANCHISE):																		
ISSUE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
OPERATIONS:																		
ROUTES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
AREAS	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES
TIME SCHEDULES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	NO	YES	NO	YES	YES
COMMODITIES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
FREIGHT RATES																		
FILED	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES	YES	YES(*)	YES	YES	YES	NO	YES	YES
REGULATED	YES	YES	NO	–	YES	NO	YES	YES	YES	NO	NO	–	YES	YES (#)	YES	NO	YES	YES

(a) POWER TO REGULATE (b) IS THIS POWER EXERCISED? FREIGHT MOTOR CARRIERS NOT REGULATED IN NEWFOUNDLAND.

(*) INTRAPROVINCIAL ONLY (#) PRESCRIBED

SOURCE: DOMINION BUREAU OF STATISTICS, SPECIAL SUPPLEMENT TO *THE MOTOR VEHICLE*, 1959, CAT. NO. 53-203 (ANNUAL), OTTAWA, OCTOBER 1960, TABLE 11.

TABLE 15

PROVINCIAL REGULATION OF TRUCKS: 1963

	PRINCE EDWARD ISL.		NOVA SCOTIA		NEW BRUNSWICK		QUEBEC		ONTARIO		MANITOBA		SASKAT-CHEWAN		ALBERTA		BRITISH COLUMBIA	
GOVERNING MOTOR CARRIER ACTS	MOTOR CARRIER ACT		MOTOR CARRIER ACT		MOTOR CARRIER ACT		TRANSP'N. BOARD ACT		PUBLIC COM. VEHICLES ACT		HIGHWAY TRAFFIC ACT		THE VEHICLES ACT		PUB. SERV. VEHICLES ACT		MOTOR CARRIER ACT	
MOTOR CARRIER REGULATORY AUTHORITIES	PUBLIC UTIL. COMM.		BOARD OF COMM'RS. OF PUB. UT.		MOTOR CARRIER BOARD		TRANSP'N. BOARD		HIGHWAY TRANSPORT BOARD		MOTOR CARRIER BOARD		HIGHWAY TRAFFIC BOARD		HIGHWAY TRAFFIC BOARD		PUBLIC UTILITY COMMISSION	
ITEMS	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
REQUIREMENT TO SHOW PUBLIC CONV. & NECESSITY	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
CARRIER OPERATING LICENSES (FRANCHISE):																		
ISSUE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
OPERATIONS:																		
ROUTES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
AREAS	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES
TIME SCHEDULES	YES	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES	YES	YES	NO	YES	NO	YES	YES
COMMODITIES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
FREIGHT RATES																		
FILED	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES(*)	YES	YES	YES	NO	YES	YES
REGULATED	YES	YES	NO	–	YES	YES	YES	YES	YES	NO	YES	YES	YES	YES (#)	YES	NO	YES	YES

(a) POWER TO REGULATE (b) IS THIS POWER EXERCISED? FREIGHT MOTOR CARRIERS NOT REGULATED IN NEWFOUNDLAND.

(*) INTRAPROVINCIAL ONLY (#) PRESCRIBED

SOURCE: DOMINION BUREAU OF STATISTICS, *THE MOTOR VEHICLE*: 1963, PART I, RATES AND REGULATIONS, CAT. NO. 53-217 (ANNUAL), OTTAWA, QUEEN'S PRINTER, MAY 1964, TABLE 4.

TABLE 16

EXCERPTS FROM ANNUAL REPORTS OF THE QUEBEC TRANSPORTATION BOARD RELATING TO TRANSPORT OF GOODS BY TRUCK

YEAR ENDING JUNE 30	1959	1960	1961	1962	1963	1964
WORK OF INVESTIGATORS NUMBER OF INSPECTIONS & INQUIRIES CONCERNING CONTROL OF TARIFFS & RATES						
QUEBEC DISTRICT	167#	79#	*	*	*	56
MONTREAL DISTRICT	185	162	115	148	115	178
DECISIONS BY ORDINANCE INVOLVING ISSUANCE, EXTENSION, MODIFICATION AND ANNULMENT OF PERMITS TO ENGAGE IN FOR-HIRE TRUCKING						
TOTAL NUMBER	1603	1579	1801	2165	2097	1871
PERCENT	100	100	100	100	100	100
NUMBER APPROVED	1349	1361	1527	1841	1473	1454
PERCENT	84.5	86.5	84.5	85.0	70.0	78.5
NUMBER REJECTED	181	171	168	135	136	91
PERCENT	11.6	10.8	9.3	6.2	6.5	4.9
NUMBER OF DIVERSE ACTIONS REQUIRING NEITHER APPROVAL NOR REJECTION	73	47	106	189	488	326
PERCENT	3.9	2.7	6.2	8.8	23.5	16.6

*NO INSPECTIONS OR INQUIRIES CONCERNING TARIFFS AND RATES REPORTED.

#NO BREAKDOWN PROVIDED BETWEEN TRUCK, BUS AND TAXI INSPECTIONS AND INQUIRIES.

TABLE 17

CORRELATION MATRIX FOR EQUATION (6) APPLICABLE ON CANADIAN INTRAPROVINCIAL FOR–HIRE TRUCKING

VARIABLE NUMBER	1	2	3	4	5	6	DESIGNATION OF VARIABLE
1	1.000	–0.654	–0.646	0.201	–0.531	–0.214	REVENUE PER TON–MILE
2		1.000	0.391	0.024	0.196	–0.175	AVERAGE LENGTH OF HAUL
3			1.000	–0.029	0.579	0.533	AVERAGE LOAD PER VEHICLE
4				1.000	–0.410	–0.506	AVERAGE FUEL TAX PER GALLON
5					1.000	0.796	AVERAGE ANNUAL LICENSE COST PER VEHICLE
6						1.000	AVERAGE ANNUAL WAGE PER EMPLOYEE

SOURCE OF DATA:. COMPUTER PRINT–OUT BASED ON EQUATION (1).

TABLE 18

CORRELATION MATRIX FOR EQUATION (8) APPLICABLE ON CANADIAN EXTRAPROVINCIAL AND INTERNATIONAL FOR–HIRE TRUCKING

VARIABLE NUMBER	1	2	3	4	5	6	DESIGNATION OF VARIABLE
1	1.000	–0.551	–0.673	0.473	–0.597	–0.655	REVENUE PER TON–MILE
2		1.000	0.366	–0.395	0.110	0.280	AVERAGE LENGTH OF HAUL
3			1.000	–0.207	0.394	0.728	AVERAGE LOAD PER VEHICLE
4				1.000	–0.410	–0.506	AVERAGE FUEL TAX PER GALLON
5					1.000	0.796	AVERAGE ANNUAL LICENSE COST PER VEHICLE
6						1.000	AVERAGE ANNUAL WAGE PER EMPLOYEE

SOURCE OF DATA: COMPUTER PRINT–OUT BASED ON EQUATION (1).

TABLE 19

CORRELATION MATRIX FOR EQUATION (9) APPLICABLE ON ALL INTERCITY TRUCKING PERFORMED BY FOR–HIRE VEHICLES OF CANADIAN REGISTRY

VARIABLE NUMBER	1	2	3	4	5	6	DESIGNATION OF VARIABLE
1	1.000	–0.648	–0.811	0.405	–0 628	–0.478	REVENUE PER TON–MILE
2		1.000	0.482	–0 170	0.253	–0.033	AVERAGE LENGTH OF HAUL
3			1 000	–0 286	0.629	0.677	AVERAGE LOAD PER VEHICLE
4				1.000	–0 410	–0.506	AVERAGE FUEL TAX PER GALLON
5					1 000	0.796	AVERAGE ANNUAL LICENSE COST PER VEHICLE
6						1.000	AVERAGE ANNUAL WAGE PER EMPLOYEE

SOURCE OF DATA. COMPUTER PRINT–OUT BASED ON EQUATION (1).

TABLE 20

CORRELATION MATRIX FOR EQUATION (10) APPLICABLE ON UNITED STATES CLASSES I AND II INTERCITY COMMON MOTOR CARRIERS

VARIABLE NUMBER	1	2	3	4	5	6	DESIGNATION OF VARIABLE
1	1 000	–0.812	–0.755	–	–0.545	–0.075	REVENUE PER TON–MILE
2		1.000	0.576	–	0.488	0.058	AVERAGE LENGTH OF HAUL
3			1.000	–	0.907	0.428	AVERAGE LOAD PER VEHICLE
4				–	–	–	AVERAGE FUEL TAX PER GALLON
5					1 000	0 471	AVERAGE ANNUAL LICENSE COST PER VEHICLE
6						1.000	AVERAGE ANNUAL WAGE PER EMPLOYEE

SOURCE OF DATA. COMPUTER PRINT–OUT BASED ON EQUATION (1).

TABLE 21

CORRELATION MATRIX FOR EQUATION (11) APPLICABLE ON ALL INTERCITY TRUCKING PERFORMED BY FOR–HIRE VEHICLES OF CANADIAN REGISTRY EXCLUDING AVERAGE FUEL TAX

VARIABLE NUMBER	1	2	3	4	5	6	DESIGNATION OF VARIABLE
1	1.000	–0.648	–0.811	–	–0.628	–0.478	REVENUE PER TON–MILE
2		1.000	0.482	–	0.253	–0.033	AVERAGE LENGTH OF HAUL
3			1.000	–	0.629	0.677	AVERAGE LOAD PER VEHICLE
4				–	–	–	AVERAGE FUEL TAX PER GALLON
5					1.000	0.796	AVERAGE ANNUAL LICENSE COST PER VEHICLE
6						1.000	AVERAGE ANNUAL WAGE PER EMPLOYEE

SOURCE OF DATA: COMPUTER PRINT–OUT BASED ON EQUATION (1).

TABLE 22

CORRELATION MATRIX FOR EQUATION (12) APPLICABLE ON UNITED STATES CLASSES I AND II INTERCITY COMMON MOTOR CARRIERS AND ALL CANADIAN FOR–HIRE INTERCITY TRUCKING COMBINED

VARIABLE NUMBER	1	2	3	4	5	6	DESIGNATION OF VARIABLE
1	1.000	–0.602	–0.784	–	–0.443	–0.191	REVENUE PER TON–MILE
2		1.000	0.496	–	0.559	0.558	AVERAGE LENGTH OF HAUL
3			1.000	–	0.637	0.373	AVERAGE LOAD PER VEHICLE
4				–	–	–	AVERAGE FUEL TAX PER GALLON
5					1.000	0.683	AVERAGE ANNUAL LICENSE COST PER VEHICLE
6						1.000	AVERAGE ANNUAL WAGE PER EMPLOYEE

SOURCE OF DATA: COMPUTER PRINT–OUT BASED ON EQUATION (1)

[illegible]rences

Alberta Province, Department of Highways. *Annual Report of the Department of Highways of the Province of Alberta, 1963–64.* Edmonton, Canada: The Queen's Printer, 1964.

2. American Trucking Associations, Inc. *Regulation or Deregulation in Public Transportation. . . . Which Way in the 1970's?* Washington, D. C.: Regular Common Carrier Conference, American Trucking Associations, Inc., 1970.
3. Burstein, M. L., Cabot, A. V., Egan, J. W., Hurter, A. P., jr., and Warner, S. L. *The Cost of Trucking: Econometric Analysis.* Dubuque, Iowa: William C. Brown Company, 1965.
4. Canada, Dominion Bureau of Statistics. *Motor Carriers—Freight.* Ottawa: The Queen's Printer, annual. Catalogue No. 53–205 through 1960; thereafter, Catalogue No. 53–222 (Class 1 and 2 carriers) and Catalogue No. 53–223 (Class 3 and 4 carriers).
5. ———, *Motor Transport Traffic.* Ottawa: The Queen's Printer, 1957–1964. Catalogue No.s 53–207 through 53–214.
6. ———. *The Motor Vehicle.* Ottawa: The Queen's Printer, annual.
7. ———. *Moving and Storage, Household Goods.* Ottawa: The Queen's Printer, annual. Catalogue No. 53–221.
8. Canadian Industrial Traffic League. *Five Traffic and Transportation Conferences: 1957–1961.* Toronto: C.I.T.L., 1961.
9. Carr, D. W. et al. "Truck-Rail Competition in Canada," In *Report of Royal Commission on Transportation,* Vol. 3. Ottawa: The Queen's Printer, 1962.
10. Commerce Clearing House. *State Motor Carrier Guide.* Chicago, Illinois: Commerce Clearing House, Inc., 1969.
11. Croxton, F. E. and Cowden, D. J. *Applied General Statistics,* 2nd ed. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1955.
12. Currie, A. W. *Canadian Transportation Economics.* Toronto: University of Toronto Press, 1967.
13. Dartnell, A. L. *The Transportation of Freight by Road in Canada.* Ph.D. Dissertation, McGill University, November 1967. Microfilm copy in the library of The Transportation Center at Northwestern University.
14. Edwards, A. L. *Statistical Analysis,* rev. ed. New York: Holt, Rinehart and Winston, 1958.
15. Fisher, R. A. *Statistical Methods for Research Workers,* 13th ed., rev. New York: Hafner Publishing Co., Inc., 1958.
16. Fraser, D. A. S. *Statistics—An Introduction.* New York: John Wiley & Sons, 1958.
17. Harper, D. V. *Economic Regulation of Motor Trucking by the States.* Illinois Studies in the Social Sciences, Vol. 43. Urbana, Illinois: The University of Illinois Press, 1959.
18. Mackey, M. C. "Remarks Prepared for Delivery before the Annual Meeting of the Transportation Committee of the National Council of Farmer Cooperatives," October 24, 1968. Mimeographed.
19. Miklius, W. *An analysis of Structure, Conduct, and Performance of the Sector of Motor Carrier Industry Operating Under Argicultural Exemption.* Ph.D. Dissertation, University of California at Los Angeles, 1964. Authorized facsimile by University Microfilms, Inc., Ann Arbor, Michigan.
20. Pegrum, D. F. *Transportation: Economics and Public Policy,* rev. ed. Homewood, Illinois: Richard D. Irwin, Inc., 1968.
21. Phillips, C. F. *The Economics of Regulation,* rev. ed. Homewood, Illinois: Richard D. Irwin, Inc., 1969.
22. Quebec Province. *Rapport Annuel de la Regie Des Transports* Quebec: The Transportation Board, annual.
23. Quebec Tariff Bureau, Inc. *Report on Agreed Charges Contracts by the Standing Rate Committee.* Montreal, Canada: 1969. Mimeographed.
24. Samuelson, P. A. *Economics: An Introductory Analysis,* 8th ed. New York: McGraw-Hill, 1970.
25. Sloss, J. "The Demand for Intercity Motor Freight Transport: A Macroeconomic Analysis," *The Journal of Business,* forthcoming.

26. Stafford, G. M. "Is Regulation No Longer Necessary?" Remarks t the Winter Meeting, Board of Governors of the Regular Common C Conference, Honolulu, Hawaii, February 20, 1970. Mimeographed.
27. Steeves, E. T. *Trucking in Canada: 1957–1967*. Paper prepared for the ann meeting of Canadian Trucking Associations, June 25–27, 1969.
28. Turgeon, W. F. A. *Report of the Royal Commission on Agreed Charges* Ottawa: The Queen's Printer, 1955.
29. U. S. Department of Agriculture. *Interstate Trucking of Fresh and Frozen Poultry under Agricultural Exemption.* Markering Research Report No. 224, Washington, D. C.: U. S. Government Printing Office, 1958.
30. U. S. Government. *Economic Report of the President Transmitted to the Congress February 1970 Together with the Annual Report of the Council of Economic Advisors.* Washington, D. C.: U. S. Government Printing Office, 1970.
31. U. S. Interstate Commerce Commission. *78th Annual Report for Fiscal Year ended June 20, 1964.* Washington, D. C.,: U. S. Government Printing Office, 1964.
32. ———. *Transport Statistics in the United States,* "Part 7, Motor Carriers." Washington, D. C.: U. S. Government Printing Office, annual.
33. U. S. Senate, Committee on Commerce. *The Motor Carrier Act of 1935: An Evaluation of the Motor Carrier Act of 1935 on the Thirtieth Anniversary of its Enactment.* Prepared under the direction of Warren G. Magnuson, Chairman, Senate Committee on Commerce, 89th Congress, 1st session, 1965.
34. Von Mises, R. *Mathematical Theory of Probability and Statistics.* New York: The Academic Press, 1964.
35. Wagner, W. H. *A Legislative History of the Motor Carrier Act, 1935.* Denton, Maryland: Roe Publishing Company, 1935.
36. Weeks, Sinclair. *Revision of Federal Transportation Policy—A Report to the President.* Washington, D. C.: U. S. Government Printing Office, 1955.
37. Welborn, D. M. *The Certification of Motor Carriers of Property: A Study in National Regulatory Administration.* Ph.D. Dissertation, University of Texas, 1962. Authorized facsimile by University Microfilms, Inc., Ann Arbor, Michigan.
38. Wilcox, C. *Public Policies Toward Business,* 3rd ed. Homewood, Illinois: Richard D. Irwin, Inc., 1966.
39. Williams, E. W. and Bluestone, D. W. *Rationale of Federal Transportation Policy.* Appendix to: U. S. Department of Commerce, *Federal Transportation Policy and Program,* Washington, D. C., U. S. Government Printing Office, 1960.